THE UNIVERSITY
of EDINBURGH

On the genetics of intermediate phenotypes and their utility

Andrew D. Bretherick

Doctor of Philosophy (PhD)

The University of Edinburgh

2020

Abstract

The vast majority of common disease-associated genetic variation is non-coding. However, the route by which non-coding genetic variation influences disease susceptibility is largely unknown. The dissection of the genetic control of variation in intermediate phenotypes, such as protein abundance or DNA methylation status, represents an important method to interrogate the pathway between genotype and phenotype.

Using array-based technologies, I assessed the genetic associations of 573,027 CpG sites in 5,101 individuals, and 249 plasma proteins in two cohorts, one of 909 and the other of 998 individuals. In addition, using mass-spectrometry, I assessed the genetic association for 4,433 proteins in the peripheral blood mononuclear cells of 251 individuals. These analyses generated a wealth of genetic associations that were further exploited in a number of ways, including Mendelian Randomisation, co-localisation with expression (RNA) quantitative trait loci (eQTLs), and enrichment analyses.

Using the genetic associations of the 249 plasma proteins, I performed proteome-by-phenome Mendelian Randomisation and demonstrated 509 putative causal links between various proteins and outcome diseases and traits, including such links to cardiovascular disease and schizophrenia. However, total plasma protein abundance derives from multiple sources and is unlikely to be representative of any single cell-type or tissue. Therefore, in an exploratory analysis, I demonstrate the feasibility of studying the cellular proteome of peripheral blood mononuclear cells using mass-spectrometry. Mass-spectrometry

proteomics provides a depth of coverage of the proteome not currently possible with other technologies, as well as enabling the possibility of additional complementary future analyses, for example, that of protein post-translational modifications.

I identified potential molecular intermediates mediating inter-chromosomal methylation quantitative trait loci (meQTLs) by assessing their co-localisation with *locally*- acting eQTLs. I found strong enrichment for genes encoding C2H2-ZF transcription factors, especially those containing a Krüppel associated box (KRAB) domain. In addition, I identified DNA methylation affected by dominance inter-chromosomal meQTL in the binding sites of many transcription factors and associated proteins.

Collectively, these analyses represent an assessment of the genetic control of plasma proteins and DNA methylation, with projection onto disease and other human traits. In addition, I lay the foundation for a much larger population-scale analysis of the cellular proteome of peripheral blood mononuclear cells to unprecedented depth: data acquisition for which is currently ongoing.

Lay Summary

How a cell controls the amount of each different protein it makes is far from fully understood. However, this is an important question because DNA variation between people that is associated with a disease is also commonly associated with protein abundance. Proteins are often viewed as the molecular machines of the cell, and different cell-types need different sets of these machines to perform their specific functions. Currently, there is great interest in understanding how DNA differences ultimately determine traits, such as height, weight, or heart disease risk. The path from genotype (the DNA of an individual) to outcome trait (e.g. disease) is often long, and over a complex regulatory landscape. This landscape is sculpted by many features, such as other proteins and chemical changes to the DNA itself (i.e. *epigenetics*).

Understanding the effects of DNA variation, particularly on proteins and their production, is important because it identifies points on the road between a DNA change and disease where it may be possible to intervene with a new medication.

I have investigated the effects of genetic differences between people on protein amount in their plasma (part of the blood), and this has allowed me to predict protein amount in other large genetic studies of various diseases. In many cases, this led to evidence that changes in the abundance of a specific protein contribute to causing disease. In addition, I explored the practicality of directly measuring protein amounts in blood cells themselves, rather than in the plasma, and present promising initial results.

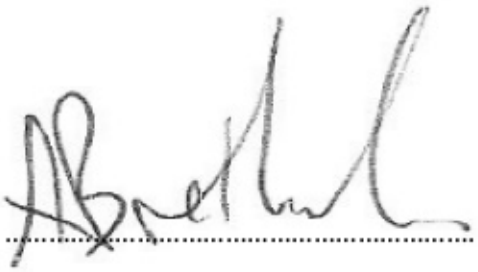
In a further study, I located where epigenetic changes are affected by DNA differences. By combining these results with those from previously published studies, I predicted many proteins likely to cause these epigenetic changes. Given the genome-wide coverage of the epigenetic data, these results provide a map of the whole landscape between genetic variation and outcome trait.

Together, these results represent an analysis of the effects of DNA variation within the cell, not only probing the underlying biology of the system, but also assisting in the identification of novel therapeutic targets.

Declarations

I declare that this thesis was written by myself and that the research reported within it is my own work, except where specified, performed under the supervision of Professor Chris Haley, Professor Chris P. Ponting, and Doctor J. Kenneth Baillie. My specific contributions to the work presented in each chapter are detailed under 'Thesis outline', below.

This work has not been previously submitted for the award of any degree at any institution.

Signed 

Date 18 May 2020

Publications

Work performed either as part of this thesis, parallel, or supplemental to it has formed part of one peer-reviewed first author publication [1], three non-first author peer-reviewed publications [2–4], one first author pre-print [5], and two non-first author pre-prints [6,7]. In addition to this, I have contributed work to the GoDMC [8] and SCALLOP [9] consortia.

The pre-print “Proteome-by-phenome Mendelian Randomisation detects 38 proteins with causal roles in human diseases and traits” [5] forms the basis of Chapter 2.

Acknowledgements

I would like to thank my supervisory team – Chris Haley, Chris Ponting, and Kenny Baillie – for their support over the course of my PhD. In addition, I would like to thank the Edinburgh Clinical Academic Track (ECAT) team (especially the directors, Jo Ness, and Rustam Al-Shahi Salman) and the Wellcome Trust for making the whole thing possible.

I would also like to thank the members of the Haley group and the wider QTL group for their day-to-day support whilst at work, and also their friendship outside of it.

Finally, I would like to thank my whole family, especially Ruth and Ada Bretherick, my parents, and my sister and her family for all their love and support throughout.

Array-based proteomics work (Chapter 2)

- I would like to thank all of the co-authors for their input to the paper.
- A debt of gratitude is owed to all the participants in all cohorts used, without whom this work would not have been possible.
- This research has been conducted using the UK Biobank Resource under project 788.
- I would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh, and the people of Orkney. DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh.

- I would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health. Genotyping was performed in the Genetics Core of the Clinical Research Facility, University of Edinburgh.

Mass Spec proteomics work (Chapter 3)

- I would like to thank all the participants of Generation Scotland.
- I would like to acknowledge the IGMM mass-spectrometry team who have been very generous with their time and have provided invaluable expertise to the project: especially Jimi Wills, and Alex von Kreigsheim.
- I would like to thank the Baillie Group, and all those at the Roslin who helped with the genesis of pilot data and attempts at cell-sorting.

meQTL work (Chapter 4)

- A vote of thanks is owed to all the participants of Generation Scotland, without whom this study would not have been possible.
- DNA methylation in Generation Scotland was obtained as part of the STRADL project who graciously allowed my use of it.

Funding

- I would like to acknowledge funding from the Wellcome PhD training fellowship for clinicians (204979/Z/16/Z), the Edinburgh Clinical Academic Track (ECAT) programme.
- The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to James F. Wilson, the MRC Human Genetics Unit quinquennial programme "QTL in Health and Disease", Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947).
- The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants (108-1080315-0302).
- Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping of the GS samples was carried out by the Genetics Core Laboratory at the Wellcome Trust Clinical Research Facility, Edinburgh, Scotland. Genotyping and DNA CpG methylation measurement was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award STratifying Resilience and Depression Longitudinally (STRADL); 104036/Z/14/Z).
- Mass-spectrometry proteomics was funded by the Wellcome Trust (204979/Z/16/Z), with supplemental funding from the MRC University Unit Programme Grants to the Human Genetics Unit (MC_UU_00007/10).

- The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in chapter 2 were obtained from the GTEx Portal on 04 Sep 2019, and Chapters 3 and 4 on the 03 Jun 2017.

Table of Contents

List of Figures

List of Tables

1. Introduction

Background

With respect to the assessment of organism level phenotypes, we are living in an era of multitudinous genome-wide association (GWA) studies. The assessment of the correlation between standard quantitative traits (e.g. height) or diseases (e.g. coronary heart disease risk), and millions of genetic variants is now routine [10]. One striking finding to come from the years of genome-wide association (GWA) study of disease associated genetic variation is that the vast majority of disease associated DNA variation is in non-coding regions. It has been estimated that approximately 93% of common disease-associated genetic variation resides in non-coding regions [11]. However, the mechanistic downstream consequences of this variation are not well established. In order to try and bridge this gap, there is a growing movement towards '-omics' studies of intermediate phenotypes [12–14]. In this context, 'intermediate' refers to any trait that could be considered to reside on the pathway between genetic variation and an outcome trait of interest (be that a disease or other 'complex' trait, or even another '-omic' trait). These studies, as far as is practical, attempt to comprehensively assess a given type of biological molecule, for example, proteomics (proteins), methylomics (DNA CpG methylation sites), and transcriptomics (RNA). These types of studies pose new challenges relating to their assessment and understanding. However, by tying a molecular and a disease / anthropometric phenotype together by their genetics enables the inference of the effects of one trait upon the other, a causal anchor if you will.

When a genetic locus is found to control the value of a quantitative trait, it is termed a quantitative trait locus (QTL). This nomenclature has been extended to include those genomic loci associated with intermediate traits, for example, genomic locations associated with DNA CpG methylation are termed meQTL, and those associated with protein abundance, pQTL. However, not all single nucleotide polymorphism within the locus are likely to be causal. For example, it is perfectly plausible that there is only one variant exerting an effect within a locus. However, others variants found within the locus, that are not directly causal, may still be found to have an association with trait. In this case, this is due to linkage disequilibrium with a causal variant.

Despite the increased complexity, the analysis of intermediate traits enables one to begin to obtain mechanistic understanding of the consequences of this non-coding variation. For example, it is estimated that about a third of disease and trait associated SNPs are also *distantly*- acting meQTL [15]. It has previously been shown that, on a small scale, this is due to the *distantly*- acting meQTL affecting the local production of a nearby transcription factor, and it is the transcription factor itself that then goes on to exert effects on distant CpGs sites [15]. This example neatly demonstrates how various intermediate traits can be interlinked and mechanistic understanding advanced.

Throughout this thesis, I have chosen to describe QTL as *locally*- or *distantly*- acting if discrimination was based upon genomic location, and *cis*- or *trans*- if the discrimination was based upon molecular mechanism (i.e. acting on the same DNA molecule or not). Note that all inter-chromosomal links would be classified as both *distant*- and to act in *trans*-.

However, whilst most *local*- variation may well act in *cis*, it does not necessarily do so. This situation becomes even more complex for *distant*- links within a single chromosome where, without additional experimentation, the molecular mechanism underlying the relationship (i.e. *cis*- vs. *trans*-) remains ambiguous.

In this thesis, I perform assessment of pQTL from plasma, cellular pQTL in peripheral blood mononuclear cells (PBMCs), and meQTL from whole blood, the backgrounds of which I briefly review here.

The human proteome

It is estimated that there are around 20,000 [16–18] genes encoded in the human genome. However, there are many more potential protein products: due to alternative splicing and post-translational modification, for example. True coverage of all protein species is technologically impossible at the present time, and is unlikely to become so in the near future. However, the term 'proteomics' tends to be used (rather loosely) for studies assessing more than a hundred or so proteins simultaneously, a convention I do not attempt to redress here.

Classically, the difficulty with traditional measurements of protein abundance, for example western blotting, is that they are not easily scalable. Current technologies used to assess the proteome at population-scales can broadly be categorised into two groups: antibody-based assays, and aptamer-based assays. Currently, to the best of my knowledge, the deepest

population-scale proteomics study available is that of Sun et al. from 2018 [14], who assessed 3,622 plasma proteins in 3,301 healthy participants. However, due to the subjective selection of proteins for inclusion on an array, array-based technologies are not a panacea for this type of work. It also remains an open question as to how representative the proteome of plasma is of any single cell-type or tissue. One might suspect, given that the blood is one of the principal transport mechanisms available to the body: responsible for oxygen delivery, metabolite transport, and the carriage of many cells in its own right – that it is unlikely to be a close proxy of the proteome of any specific cell-type or tissue. Due to the multiple origins of proteins in the plasma, proteins that are expressed in a limited subset of cell-types are likely to be diluted when added to the milieu of the plasma, further compounding difficulties in their detection.

An alternative approach is that of mass-spectrometry. When using cell-lysate, the depth of coverage possible with high-performance liquid-chromatography mass-spectrometry is unparalleled by current array-based technologies. This technology is now mature enough to begin to apply at population-scales and some early mass-spectrometry-based genome-wide association studies are beginning to emerge: for example, in induced pluripotent stem cells, Mirauta et al. [19] revealed over 700 human pQTL in induced pluripotent stem cells (false discovery rate (FDR) <0.10).

The human methylome

Cytosine methylation of DNA is an important epigenetic modification in Eukaryotes and, when disturbed, has been associated with numerous diseases (be that as cause or consequence) [12,20,21]. In molecular terms, it is the covalent addition of a methyl group to the DNA base cytosine, resulting in 5-methylcytosine. In humans, there is generally a paucity of CpG dinucleotides out with of 'islands' of relatively high CpG content, often associated with gene promoters [22]. It has been shown that, in humans, 5-methylcytosine predominantly occurs at CpG dinucleotides and, outside of CpG islands (where the situation is more complex), is wide-spread across the genome [23].

Recent technological advances have enabled the production of array-based technologies to assess the DNA CpG methylation status at >500,000 sites simultaneously. Indeed, previous smaller GWA studies have been performed on DNA methylation [12,21], with some degree of success. For example, it has previously been demonstrated that the phenotypic variation explained by meQTL is substantial. In one study, 8% of *locally*- acting meQTL explained more than 50% of the variation of the CpG to which they corresponded [12]. In that same study, despite fewer *distantly*- acting meQTL explaining very large proportions of CpG variance when compared to *locally*- acting meQTL, an enrichment of *distantly*- acting meQTL was still found in upstream and 5' untranslated regions of genes.

Genome-wide association

Standard GWA studies commonly focus on the estimation of allelic effects, that is, the effect per additional encoded allele. Using this model, these studies have been fantastically successful at identifying disease and trait associated genetic variation. However, this is not the only model it is possible to fit. For example, in chapter 4, I fit a model that explicitly includes both an additive and dominance effect of the encoded allele. Given the increased complexity of the model as well as difficulties of detecting dominance variation (as discussed in the chapter), molecular traits represent a bastion where fitting this model may be useful, in particular, given the predicted, bi-allelic effects of diffusible mediators.

What is the utility of these genetic associations?

Co-localisation of effects at a given locus can give information about the likely molecular actors orchestrating the relationship between genotype and phenotype. For example, IL6 receptor signalling and coronary heart disease [24,25]. As mentioned above, methods exist – such as Mendelian Randomisation (MR) [26,27] – that can combine GWAS and '-omics' to infer disease risk-altering molecules. It is worth remembering that molecules identified as causally contributing to disease risk are potential drug targets, and their identification may have profound effects on the therapeutics of the future. It is estimated that genetic evidence in support of a protein's candidature as a drug target could double its probability of success in clinical development [28].

Genetic differences between individuals can affect the abundance of molecular intermediates: including protein (pQTL), and DNA CpG methylation (meQTL). The phrase 'genetic control' is used within this thesis to mean this type of direct link between genotype and phenotype. Despite leveraging *a priori* biological knowledge as the justification for doing this, it has often not directly been observed, and therefore not absolute: this should be considered wherever this phrase is used. Given this caveat, if an allele causes a change in the mean abundance of a protein then, within the assumptions of MR (discussed in Chapters 2 and 5), this implies that any observed association of the allele with disease risk is due to differences in mean protein level: that is, the protein causally contributes to disease risk. In essence, MR is akin to a naturally occurring randomised control trial. For MR, local protein quantitative trait loci (pQTLs) are almost ideal instrumental variables: they have large effect-sizes and highly plausible biological relationships with protein level and, when used, may provide quantitative information on directly druggable protein targets.

Thesis outline

Chapter 1: Introduction

Chapter 2: Plasma array-based proteomics

In this chapter, I present an unpublished paper entitled "Linking protein to phenotype with Mendelian Randomisation detects 38 proteins with causal roles in human diseases and traits" [5]. Author contributions are included within the paper itself; however, to summarise: I performed the GWA of the proteins, in both cohorts, and downstream analyses (other than the ChEMBL search, and the composition of Figure2d).

In the manuscript I present proteome-by-phenome MR as a paradigm for large-scale drug target discovery. This is a topic of great interest to both academia and the pharmaceutical industry. Using proteome-by-phenome MR, I report 509 robust protein-outcome links, each likely to causally contribute to the outcome trait. These include: FABP2 in cardio-vascular disease, SHPS1 in schizophrenia, and IL6R in atopy. This study and its results are analogous to 509 randomised control trials, in humans, in which each reports a significant finding.

In order to perform proteome-by-phenome MR I undertook genome-wide association with plasma concentration of 249 proteins in two European cohorts: each of >850 individuals, discovery and replication. I used outcome data on 846 traits and diseases, from published studies of UK Biobank [29] and others (see Chapter), to assess 54,144 exposure-outcome pairs. I identified 38 proteins inferred to causally contribute to 509 significant exposure-outcome pairs and, for significant outcomes from UK Biobank, I also tested whether results had consistent MR effect estimates across each locus, or not.

Proteome guided MR reveals novel candidate targets for drug discovery and offers insights into the likely side-effects of therapeutic agents that target a particular molecular entity. This approach is less expensive, less time-consuming, and more scalable than randomised control trials, and more physiologically relevant than model organism studies.

Chapter 3: Cellular mass-spectrometry based proteomics

Here I present the results of the first phase of a much larger project to assess the proteome of peripheral blood mononuclear cells at both unprecedented depth, and scale in Generation Scotland. I conceived, designed, and executed this project with the help and support of the IGMM mass-spectrometry team, Generation Scotland, and my supervisory team. Generation Scotland had already collected the PBMCs and undertaken genotyping and associated QC.

Ultimately, this project will include the proteomes of the peripheral blood mononuclear cells of over 850 individuals: in the chapter I present the results from the first phase of 251 individuals. I identify 4,433 proteins, each found in one hundred or more samples, at a false discovery rate of less than one percent. To the best of my knowledge, this is already the deepest assessment of the proteome of peripheral blood mononuclear cells, and the broadest population-scale assessment of the proteome of any primary cell-type. The direction and magnitude of effect of my results are broadly consistent with those obtained for RNA in whole blood, yet have both incomplete overlap, and imperfect correlation. Comparison to plasma protein quantitative trait loci hints at unique genomic signatures of control, commensurate with the differing origins of the proteins in peripheral blood mononuclear cells and the plasma.

I assess those proteins for which I identified a significant (FDR <0.05) *locally*- acting pQTL (n = 108) for causal contribution to rheumatoid arthritis risk and identify 7 proteins potentially

involved in its pathogenesis (HLA-DQA1, NELFE, PADI4, HLA-B, RNASET2, PADI2, and HLA-F), including both members (PADI2 and PADI4) of the 'histone H3-R26 citrullination' pathway (GO:0036413). In reviewing the association of the lead-SNP at the *locally*- acting pQTL associated with the proteins PADI2 and PADI4 across their respective genes, I demonstrate an association of the lead-pQTL SNP with DNA methylation across the gene of the protein to which the SNP is associated, representing a potential method to assess for potential pleiotropy.

Chapter 4: DNA CpG methylation

I perform a genome-wide association study of >500,000 CpG sites, using the Illumina EPIC array, in 5,101 individuals from Generation Scotland and demonstrated that the relationship between transcription factors and *inter-chromosomal meQTLs* is very strong.

Using *locally*- acting expression quantitative trait loci (eQTL) from the Genotype-Tissue Expression (GTEx) project (v6p) [13] to identify genes locally impacted by the SNPs of inter-chromosomal meQTLs included 5.8% of all known human transcription factors: including 9.7% of all known human C2H2 zinc-finger (C2H2-ZF) transcription factors, and 14.9% of all known human Krüppel associated-box (KRAB) containing transcription factors. This is a significant enrichment when compared to a random background of eQTL from the GTEx (v6p) [13] project. In addition, the proportion of C2H2 and KRAB containing transcription factors in the set of transcription factors identified is significantly enriched when compared to all known human transcription factors [30].

The KRAB domain is classically considered a transcriptional repressor and it has previously been shown that the DNA methylation induced by KRAB domain containing transcription factors can spread from their binding sites [31,32]. Consistent with this, I demonstrate an association with increased dominance inter-chromosomal genetic control of DNA CpG methylation for CpGs located in the binding sites of many transcription factors, DNA binding proteins, and their co-factors: not limited to those I identify as directly associated with inter-chromosomal meQTL.

Generation Scotland and STRADL had already performed genotyping and imputation, and the genesis of the DNA CpG methylation data. Quality control and residualisation of the methylation data was performed by Rosie Walker and Yanni Zeng. The text describing the genotyping and measurement of methylation subsections of the Methods section of Chapter 4 is partially shared with the GoDMC manuscript. I created the additive / dominance GWA program and pipeline, and performed all downstream analyses of these results.

Chapter 5: Conclusion

Finally, I close with a discussion of the foregoing and a forward-looking section regarding future possibilities.

2. Plasma array-based proteomics

Pre-material

This chapter comprises an edited version of a revised pre-publication paper: a pre-print of which is available on BioRxiv: https://www.biorxiv.org/content/10.1101/631747v1.

Paper: Linking protein to phenotype with Mendelian Randomisation detects 38 proteins with causal roles in human diseases and traits

ANDREW D. BRETHERICK[1*], ORIOL CANELA-XANDRI[1,2], PETER K. JOSHI[3], DAVID W. CLARK[3], KONRAD RAWLIK[2], THIBAUD S. BOUTIN[1], YANNI ZENG[1,4,5,6], CARMEN AMADOR[1], PAU NAVARRO[1], IGOR RUDAN[3], ALAN F. WRIGHT[1], HARRY CAMPBELL[3], VERONIQUE VITART[1], CAROLINE HAYWARD[1], JAMES F. WILSON[1,3], ALBERT TENESA[1,2], CHRIS P. PONTING[1], J. KENNETH BAILLIE[2], AND CHRIS HALEY[1,2*]

[1] *MRC Human Genetics Unit, Institute of Genetics and Molecular Medicine, University of Edinburgh, Western General Hospital, Crewe Road, EH4 2XU, Scotland, UK.*

[2] *The Roslin Institute, University of Edinburgh, Easter Bush, EH25 9RG, Scotland, UK.*

[3] *Centre for Global Health Research, Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Teviot Place, Edinburgh, EH8 9AG, Scotland, UK.*

[4] *Faculty of Forensic Medicine, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

[5] *Guangdong Province Translational Forensic Medicine Engineering Technology Research Center, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

[6] *Guangdong Province Key Laboratory of Brain Function and Disease, Zhongshan School of Medicine, Sun Yat-Sen University, 74 Zhongshan 2nd Road, Guangzhou 510080, China.*

** Corresponding Author*

Abstract. To efficiently transform genetic associations into drug targets requires evidence that a particular gene, and its encoded protein, contribute causally to a disease. To achieve this, we employ a three-step proteome-by-phenome Mendelian Randomisation (MR) approach. In step one, 154 protein quantitative trait loci (pQTLs) were identified and independently replicated. From these pQTLs, 64 replicated locally-acting variants were used as instrumental variables for proteome-by-phenome MR across 846 traits (step two). When its assumptions are met, proteome-by-phenome MR, is equivalent to simultaneously running many randomised controlled trials. Step 2 yielded 38 proteins that significantly predicted variation in traits and diseases in 509 instances. Amongst the 271 instances from GeneAtlas (UK Biobank), 77 showed little evidence of pleiotropy (step three). Results were wide ranging: including, for example, new evidence for a causal role of tyrosine-protein phosphatase non-receptor type substrate 1 (SHPS1; *SIRPA*) in schizophrenia, and a new finding that intestinal fatty acid binding protein (FABP2) abundance contributes to the pathogenesis of cardiovascular disease. We also demonstrated confirmatory evidence for the causal role of four further proteins (FGF5, IL6R, LPL, LTA) in cardiovascular disease risk.

Author summary. The targets of most medications prescribed today are proteins. For many common diseases our understanding of the underlying causes is often incomplete, and our ability to predict whether new drugs will be effective is remarkably poor. Attempts to use genetics to identify drug targets have an important limitation: standard study designs link disease risk to DNA but do not explain how the genotype leads to disease. In our study, we made robust statistical links between DNA variants and blood levels of 249 proteins, in two separate groups of Europeans. We then used this information to predict protein levels in large genetic studies. In many cases, this second step gives us evidence that high or low levels of a given protein play a role in causing a given disease. Among dozens of high-confidence links, we found new evidence for a causal role of a protein called SHPS1 in schizophrenia, and of another protein (FABP2) in heart disease. Our method takes advantage of information from large numbers of existing genetic studies to prioritise specific proteins as drug targets.

Introduction

An initial goal of drug development is the identification of targets – in most cases, proteins – whose interaction with a drug ameliorates the development, progression, or symptoms of disease. After some success, the rate of discovery of new targets has not accelerated despite substantially increased investment[33]. A large proportion of drugs fail during the last stages of development – clinical trials – because their targets do not alter whole-organism phenotypes as expected from observational and other pre-clinical research[34]. Genetic approaches to drug development[35] offer a distinct advantage over observational studies. It is estimated that by selecting targets with genetic evidence, the chance of success of those targets doubles in subsequent clinical development[28]. For example, a recent study found that 12% of all targets for licenced drugs could be rediscovered using GWA studies[36]. Indeed, there have been a number of recent high-profile successes prioritising therapeutic targets at genome-wide scales[37,38]. Nevertheless, the genetic associations of disease are often still not immediately interpretable[39] and many disease-associated variants alter protein levels via poorly understood mechanisms.

When combined with proteomic data, however, genetics can provide insight into proteins that likely impact disease pathogenesis. Mendelian Randomisation (MR) in this context uses genetic variants to estimate the effect of an exposure on an outcome, using the randomness by which alleles are allocated to gametes to remove the effects of unmeasured confounding between a protein and the outcome[40]. Given a set of assumptions, detailed below, this approach is analogous to a naturally-occurring randomised controlled trial. Using a genetic variant that predicts the abundance of a mediating molecule, MR tests the hypothesis that

this molecule plays a causal role in disease risk. To do so it takes advantage of the patient's, or participant's, randomisation at conception to this molecule's genetically-determined level. Under this model, it is possible to use population level genetic information to draw causal inference from observational data.

Proteome-by-phenome MR, in common with all other MR studies, has three key assumptions that must be fulfilled to ensure the legitimacy of any causal conclusions drawn[26]: 1) that the SNP is associated with the exposure of interest, 2) that the SNP is independent of any confounders, and 3) that the SNP does not influence the outcome of interest, except via the exposure variable.

A common concern in the use of MR is that the genetic variant is linked to the outcome phenotype via an alternative causal pathway. In a drug trial this would be analogous to an intervention influencing a clinical outcome through a different pathway than via its reported target. To avoid pursuing drugs that target an irrelevant molecular entity, and hence that have no beneficial effect, we applied MR to proteins – the likely targets of therapy – and limited our genetic variants to those that are locally-acting protein quantitative trait loci (pQTLs). This approach provides stronger supporting evidence for a causal role of the protein on disease than relying on the proximity of a disease-associated genetic variant to a nearby gene, or using mRNA abundance as a proxy for protein abundance[19].

Previous studies have also leveraged the increased availability of pQTL data for drug target and biomarker discovery[14,41–46]. For example, in one of the largest pQTL studies to date, Sun et al.[14] applied an aptamer-based approach (rather than an antibody-based assay as here) to perform extensive co-localisation analyses and used MR to assess the causal contribution of IL1RL1–IL18R1 locus to atopic dermatitis, and that of MMP12 to coronary heart disease. In the study presented here, we attempt to systematically use MR to link protein to outcome trait by taking a three-step approach. Firstly, identifying replicated pQTL in our two European cohort studies before then using these in a systematic MR approach with two large sets of GWA study data. In a final step, we test results from one of these sets for their consistency with a single underlying causal variant (affecting both variation in protein concentration and outcome phenotype).

Overall, our proteome-by-phenome MR approach assessed the causal role of 64 proteins in 846 outcomes (e.g. diseases, anthropomorphic measures, etc.), identifying 38 as causally contributing to human diseases or other quantitative traits. Notwithstanding the assumptions of MR, obtaining evidence for causality from studies such as this is far more scalable than via randomised controlled trials, and is more physiologically relevant than model organism studies.

Results

Protein QTLs

The abundance of an individual protein can be associated with DNA variants that are either local or distant to its gene (termed local- and distal-pQTLs, respectively). In many respects,

locally-acting pQTLs are ideal instrumental variables for MR: they tend to have large effect sizes, have highly plausible biological relationships with protein level, and provide quantitative information about (often) directly druggable protein targets. This is in contrast to distal pQTLs, where the pathway through which they exert their effects is generally unknown, with no *a priori* expectation of a direct effect on a single target gene.

We assayed the plasma levels of 249 proteins using high-throughput, multiplex immunoassays and then performed genome-wide association of these levels in each of two independent cohorts (discovery and replication) of 909 and 998 European individuals who had previously been genotyped.

Lead-SNPs, defined as the variant with the smallest p-value and accounting for linkage disequilibrium (Methods), were identified for each protein. As expected, pQTLs were highly concordant between the two independent cohorts (Supplementary Table 1). 121 pQTL were identified in the discovery dataset, and, of these, 90.1% (109/121) were successfully replicated after accounting for multiple testing in both the discovery and replication. However, this was felt to be excessively stringent with respect to instrument identification, and a more permissive threshold of $5x10^{-8}$ was therefore used in the discovery cohort. Of the 209 lead-SNPs identified in the discovery cohort at this threshold, 154 were successfully replicated (accounting for multiple testing during replication and with consistent direction of effect). These represented pQTLs for 82 proteins, all but two proteins were successfully mapped to an autosomal gene (Ensembl GRCh37). The majority of these proteins (64/80; 80%) had a replicated lead-SNP within 150kb of the gene encoding the protein (Figure 1).

The variant to use as the instrumental variable for each protein was selected as the replicated lead-SNP lying within 150kb of the gene encoding the protein with the lowest significant p-value in the discovery set (Methods). Increasing this proximity threshold to within 1Mb added a single protein only. Further support for the validity of these instruments was provided through comparison with the results of Sun et al.[14] and GTEx[13] (Methods): of the instrumental variables identified (a) 52% (14/27) of those comparable were in high LD (r^2>0.8) with the results of Sun et al. (Supplementary Table 2), and (b) 30% (16/54) were also called as significant expression QTLs (eQTLs; Bonferroni correction; Supplementary Table 3) in GTEx – in keeping with previous studies[14].

Proteome-by-phenome Mendelian Randomisation

Proteome-by-phenome MR was then applied to 54,144 protein-trait pairs obtained from these 64 replicated local-pQTLs and 778 traits obtained from GeneAtlas (UK Biobank)[29], and 68 traits from 20 additional genome-wide association (meta-analysis) studies[47–66] identified through Phenoscanner[67,68] (Figure 1; Supplementary Table 4; Methods). Phenoscanner studies were additionally analysed because, although the UK Biobank cohort is large (~500,000 individuals), for many diseases the number of affected individuals is small, resulting in low statistical power (Methods).

Proteome-by-phenome MR yielded 271 significant protein-trait pairs (FDR <0.05) in GeneAtlas, and 238 significant (FDR <0.05) pairs using Phenoscanner data. Thirty-two of the 64 proteins were causally implicated for one or more traits in GeneAtlas, and 36 of 64 in the

Phenoscanner studies' traits. GeneAtlas and Phenoscanner traits are not mutually exclusive, and some of the Phenoscanner studies included UK Biobank data. Nevertheless, a majority (60%; 38/64) of the proteins were implicated in one or more traits (e.g. IL6R: as discussed below; Supplementary Table 5 and Supplementary Table 6).

For some of these inferences, genetic evidence of an association between a protein and phenotype has previously been proposed based simply on physical proximity of the genes to GWA intervals. However, in actually measuring protein products we go well beyond genetic proximity-based annotation of GWA hits: (a) we provide direct evidence that a SNP actually changes the abundance of a protein, and (b) notwithstanding the assumptions of MR, that the change in protein abundance observed is consistent with a causal effect of the protein on outcome trait variation. In addition, notwithstanding the different significance criteria, nearly two-thirds (62%; 318/509) of the significant (FDR <0.05) MR associations between protein and outcome were not matched by significant (p-value $<5x10^{-8}$) association of the DNA variant to outcome.

Heterogeneity of effect-size estimates

For GeneAtlas results, we use HEIDI to test for heterogeneity of MR effect estimates between the lead variant (the primary instrument) and those of linked variants. More specifically, this method tests the null hypothesis that the observed MR result is consistent with a single causal variant[27], explicitly accounting for the LD structure across the locus. In these results, 77 of 271 survived the HEIDI heterogeneity test (p-value >0.05). These 77

proteins thus have: (1) high-quality evidence of association to a DNA variant that provides congruent predictions for both plasma protein levels and disease risk or trait, and (2) a low risk of pleiotropy, because of both the physical proximity of the pQTL to the protein's gene, and their survival of the HEIDI test (Supplementary Table 5). These 77 relationships provide the most robust evidence that the level of the protein directly alters disease risk or trait.

Nevertheless, we emphasise that all 509 causal inferences (271 from GeneAtlas[29] and 238 from studies identified through Phenoscanner[67,68]; Figure 2, and Supplementary Table 5 and Supplementary Table 6), even those consistent with heterogeneity (GeneAtlas only), remain potential high quality drug targets. An appropriate interpretation of this result is that there are 271 potentially causal links identified in GeneAtlas, with additional support for 77 based on results of the HEIDI analysis. This is because the HEIDI heterogeneity test (Figure 1) is susceptible to type I errors (i.e. false positives) in the context of this study. The method can report significant heterogeneity where there is, in fact, none if: (a) there are multiple causal variants present within a locus, or (b) there are differences in the LD structure among the discovery pQTL GWA population (used for lead-SNP selection), the replication pQTL GWA study population (used for effect-size estimation), the outcome trait GWA study population, or that of the LD reference. Finally, it is worth noting that we applied the HEIDI test in a conservative manner: a significant HEIDI test implies heterogeneity yet we did not apply a multiple testing correction. Applying a Bonferroni correction (271 tests) to the HEIDI p-value, yields 180 of the protein-outcome pairs (rather than 77) as not significantly heterogeneous.

Tractability of the proteins assessed as therapeutic targets

Of the 32 proteins for which we identified a significant MR association in GeneAtlas (Supplementary Table 5), we found 1,319 compounds (Supplementary Table 7) associated with 10 proteins in ChEMBL. Of these compounds, 10 have already been tested in phase 2, or greater, trials: targeting DLK1, LPL, and LGALS3.

Our results draw causal inference between the plasma concentration of specific proteins and many diseases and outcome phenotypes. For example, we provide supporting evidence for a role of IL4R in asthma, IL2RA in thyroid dysfunction, and IL12B in psoriasis (Figure 2), as well as many cellular phenotypes, such as Transferrin receptor protein 1 (encoded by *TFRC*) in mean corpuscular haemoglobin. Multiple disease endpoints exist to which we have found a MR link and, additionally, for some diseases we have causal links from multiple proteins (Figure 2a, b; Supplementary Table 5 and Supplementary Table 6).

Many-to-One: multiple proteins link to asthma.

Asthma is an inflammatory condition affecting the airways. Using GeneAtlas data, our analysis finds 5 proteins – all interleukin receptors – whose levels causally contribute to asthma disease risk: IL1RL1, IL1RL2, IL2RA, IL4R, and IL6R (Figure 2d). Prior links between these proteins and asthma or atopy exist (IL1RL1[69,70] and IL1RL2[14], IL2RA[64,71], IL4R[72], and IL6R[64,72–76]), albeit not necessarily strong evidence for a causal link. Of these, IL6R was not significantly heterogeneous in HEIDI testing ($p > 0.05$), and also IL4R if accounting for multiple tests ($p > 0.05/271$). Given the association between eosinophils and asthma, it is

worth noting that IL1RL1, IL1RL2, IL2RA, and IL4R are all linked to 'Eosinophil count' and 'Eosinophil percentage' in GeneAtlas. Whilst not a true replication, due to the use of UK Biobank data in both GeneAtlas and some of the Phenoscanner studies, Figure 2d reveals strong concordance between the MR links identified between the two. Of the 12 Phenoscanner studies reporting significant MR links in this study[47,49–51,53,55,57,60,61,64–66], 5 include UK Biobank data from ~150,000 individuals[49,55,57,65,66], and only one uses the full UK Biobank release[61].

One-to-Many: Linking IL6R levels to atopy, rheumatoid arthritis, and coronary artery disease.

We also found evidence for a causal association between plasma IL6R abundance and coronary artery disease (CAD), atopy, and rheumatoid arthritis (Figure 2, Supplementary Table 5, and Supplementary Table 6). We note previous support for these inferences: for example, tocilizumab (a humanized monoclonal antibody against IL6R protein) is in clinical use for treating rheumatoid arthritis[77], prior MR evidence has linked elevated levels of soluble IL6R to reduced cardiovascular disease[24,25], and, as discussed above, there is previous genetic evidence of a link between IL6R and atopy[64,72–76].

SHPS1 and schizophrenia

Three proteins were implicated in the pathogenesis of schizophrenia: (i) Tyrosine-protein phosphatase non-receptor type substrate 1 (SHPS1; *SIRPA*) – Figure 3, (ii) Tumour necrosis factor receptor superfamily member 5 (*CD40*), and (iii) Low affinity immunoglobulin gamma Fc region receptor II-b (*FCGR2B*).

Focussing on SHPS1, it is highly expressed in the brain, especially in the neuropil (a dense network of axons, dendrites, and microglial cell processes) in the cerebral cortex (https://v18.proteinatlas.org/ENSG00000198053-SIRPA/tissue [78–80]; accessed 01 Apr 2019), and co-localises with CD47 at dendrite-axon contacts[81]. Mouse models in which the *SHPS1* gene is disrupted exhibit many nervous system abnormalities, such as reduced long term potentiation, abnormal synapse morphology and abnormal excitatory postsynaptic potential (MGI: 5558020 [82]; http://www.informatics.jax.org/; v6.13; accessed 01 Apr 2019). Other mouse and rat models link CD47 to sensorimotor gating and social behaviour phenotypes[83–87]. In addition, SHPS1 mediates activity-dependent synapse maturation[82] and may also have a role as a "don't eat me" signal to microglia[88]. SHPS1 levels tend to be lower in the dorsolateral prefrontal cortex of schizophrenia patients[89]. Finally, the observed effect of SHSP1 on schizophrenia was not significantly heterogeneous in the results of the Schizophrenia Working Group of the Psychiatric Genomics Consortium (2014) (p-value 0.53).

FABP2 and coronary artery disease

Four other proteins, in addition to IL6R, were identified as contributing to CAD pathogenesis, namely FABP2, FGF5, LPL, and LTA (Figure 2). FGF5, LPL, IL6R, and LTA had been implicated previously[49,90,91], whereas FABP2 had more limited prior evidence for its involvement.

pQTL analysis identified two lead DNA variants in close proximity (<150kb) to the *FABP2* gene. Using SNP rs17009129, we find a causal link between FABP2 abundance and CAD (p-value 1.1×10^{-4}; FDR <0.05; β_{MR} -0.11; se_{MR} 0.028; β_{MR} and se_{MR} units: log(OR)/standard

deviation of residualised protein concentration) without significant heterogeneity (p-value 0.24) which suggests shared causal genetic control. Furthermore, a second independent SNP (LD r^2 <0.2; rs6857105) replicates this observation (MR p-value $5.0x10^{-4}$; HEIDI p-value 0.34; β_{MR} -0.17; se_{MR} 0.047). Both SNPs (rs17009129, and rs6857105) fell below genome-wide significance (p-value $<5x10^{-8}$) in the full meta-analysis of van der Harst[61] on CAD. Consequently, this is the first time, to our knowledge, that variants associate with *FABP2* abundance have been demonstrated to contribute causally to CAD pathogenesis.

Discussion

Proteome-by-phenome MR efficiently and robustly yields evidence for proteins as drug targets. It offers a data-driven approach to drug discovery using population-level data, and quantifies the strength of evidence for causation. Previous studies have made successful forays into the use of pQTL in mapping protein variation onto disease[14,41–46], and both the coverage of the proteome and the availability of disease and trait GWA study results are ever increasing. By using the lead variants of locally-acting pQTLs as instrumental variables, we focused specifically on a subset of functionally relevant variants for those proteins under study: this choice reduced the multiple testing burden when compared to genome-wide scans for associations of the outcome trait.

A potential problem with antibody- and aptamer-based assays is that any perturbation to binding, such as a change to an epitope, appears incorrectly as a change in abundance. In the absence of a well-defined reference, we cannot exclude the possibility that some of the

pQTL we have called indicate epitope changes rather than changes in protein abundance. However, in each case, a bona fide biological association does exist between the genetic variant and the protein. With respect to MR, this would change the biological interpretation of the exposure only: protein abundance or sequence isoform, for example.

In addition, proteome-by-phenome MR has inherent limitations. First, a true positive MR association in our analysis implies that any intervention to replicate the effect of a given genotype would alter the relevant phenotype. Nevertheless, this association is informative neither of the time interval, during development for example, nor the anatomical location in which an intervention would need to be delivered. Second, pleiotropic effects cannot be excluded entirely without (unachievable) quantification of every mediator. Third, the abundance of a protein in plasma may be an imperfect proxy for the effect of a drug targeting that protein at the level of a whole organism. Finally, plasma abundance does not necessarily reflect activity. For example, a variant may cause expression of high levels of an inactive form of a protein. Or, for proteins with both membrane-bound and unbound forms, the MR direction of effect observed from quantifying soluble protein abundance may not reflect that of membrane-bound protein. For many membrane-bound proteins, a soluble (often antagonistic) form exists that is commonly produced through alternative splicing or proteolytic cleavage of the membrane-bound form. Based on 1,000 Genomes[92,93] data, the variant we use to predict IL6R level, rs61812598, for example, is in complete LD with the missense variant rs2228145 whose effects on proteolytic cleavage of the membrane-bound form and alternative splicing have been examined in detail[94]. Carriers of the 358Ala allele at rs2228145 tend to have increased soluble IL6R but reduced membrane-bound IL6R in a number of immune cell types. Differences between the effects of soluble and membrane-

bound forms of a protein may be widespread. For example, dupilumab is a monoclonal antibody that targets IL4R, a key component of both IL4 and IL13 signalling. It is currently under investigation for the treatment of asthma and has shown promising results in both eosinophilic and non-eosinophilic asthma[95,96]. Based on our results, we would have predicted that increased levels of IL4R result in a lower risk of asthma (Supplementary Table 5). This is in contrast to the direction-of-effect due to dupilumab administration. However, as with IL6R, IL4R has both a soluble and a membrane-bound form. Encouragingly, despite this, a relationship between dupilumab and asthma remains plausible – as evidenced by the 14 recently completed or ongoing clinical trials to assess the efficacy and safety of dupilumab in asthma (as of 26 March 2019, ClinicalTrials.gov).

As well as its utility in identifying potential therapeutic targets for drug development, proteome-by-phenome MR also allows for an assessment of potential off-target effects of existing pharmacological targets. For example, we predict an effect of IL4R modulation on eosinophil count and percentage. This is an association already realised in one of the phase II clinical trials investigating dupilumab in asthma: a rise in eosinophil count was observed for some patients, even leading to the withdrawal of one patient from the study[95].

Conclusions

In summary, we have identified dozens of plausible causal links by conducting GWA of 249 proteins, followed by phenome-wide MR using replicated locally-acting pQTLs of 64 proteins. The approach is statistically robust, relatively inexpensive, and high-throughput.

54,144 protein-outcome links were assessed and 509 significant (FDR <0.05) links identified: including anthropometric measures, haematological parameters, and diseases. Opportunities to discover larger sets of plausible causal links will increase as study sizes and pQTL numbers grow. Indeed, whole-proteome versus Biobank GWA Atlas studies will likely become feasible as pQTL measurement technologies mature further.

Methods

Cohort description. From the islands of Orkney (Scotland) and Vis (Croatia) respectively, the ORCADES[97] and CROATIA-Vis[98,99] studies are of two isolated population cohorts that are both genotyped and richly phenotyped.

The Orkney Complex Disease Study (ORCADES) is a family-based, cross-sectional study that seeks to identify genetic factors influencing cardiovascular and other disease risk in the isolated archipelago of the Orkney Isles in northern Scotland[97]. Genetic diversity in this population is decreased compared to Mainland Scotland, consistent with the high levels of endogamy historically. 2,078 participants aged 16-100 years were recruited between 2005 and 2011, most having three or four grandparents from Orkney, the remainder with two Orcadian grandparents. Fasting blood samples were collected and many health-related phenotypes and environmental exposures were measured in each individual. All participants gave written informed consent and the study was approved by Research Ethics Committees in Orkney and Aberdeen (North of Scotland REC, 26/11/2003).

The CROATIA-Vis study includes 1,008 Croatians, aged 18-93 years, who were recruited from the villages of Vis and Komiza on the Dalmatian island of Vis during spring of 2003 and 2004. All participants were volunteers and gave written informed consent. They underwent a

medical examination and interview, led by research teams from the Institute for Anthropological Research and the Andrija Stampar School of Public Health, (Zagreb, Croatia). All subjects visited the clinical research centre in the region, where they were examined in person and where fasting blood was drawn and stored for future analyses. Many biochemical and physiological measurements were performed, and questionnaires of medical history as well as lifestyle and environmental exposures were collected. The study received approval from the relevant ethics committees in Scotland (South East Scotland Research Ethics Committee, REC reference: 11/AL/0222) and Croatia (University of Split School of Medicine Ethics committee, Class:003-08/11-03/-005 No.: 2181-198-03-04/10-11-0008).

Both studies complied with the tenets of the Declaration of Helsinki.

Genotyping. Chromosomes and positions reported in this paper are from GRCh37 throughout. Genotyping of the ORCADES cohort was performed on the Illumina Human Hap 300v2, Illumina Omni Express, and Illumina Omni 1 arrays; that of the CROATIA-Vis cohort used the Illumina HumanHap300v1 array.

The genotyping array data were subject to the following quality control thresholds: genotype call-rate 0.98, per-individual call-rate 0.97, failed Hardy-Weinberg test at p-value $<1x10^{-6}$, and minor allele frequency 0.01; genomic relationship matrix and principal components were calculated using GenABEL (1.8-0)[100] and PLINK v1.90[101,102].

Assessment for ancestry outliers was performed by anchored PCA analysis when compared to all non-European populations from the 1,000 Genomes project[92,93]. Individuals with a mean-squared distance of >10% in the first two principal components were removed. Genotypes were phased using Shapeit v2.r873 and duoHMM[103] and imputed to the HRC.r1-1 reference panel[104]. 278,618 markers (Hap300) and 599,638 markers (Omni) were used for the imputation in ORCADES, and 272,930 markers for CROATIA-Vis.

Proteomics. Plasma abundance of 249 proteins was measured in two European cohorts using Olink Proseek Multiplex CVD2, CVD3, and INF panels. All proteomics measurements were obtained from fasting EDTA plasma samples. Following quality control, there were 971 individuals in ORCADES, and 887 individuals in CROATIA-Vis, who had genotype and proteomic data from Olink CVD2, 993 and 899 from Olink CVD3, and 982 and 894 from Olink INF. The Olink Proseek Multiplex method uses a matched pair of antibodies for each protein, linked to paired oligonucleotides. Binding of the antibodies to the protein brings the oligonucleotides into close proximity and permits hybridization. Following binding and extension, these oligonucleotides form the basis of a quantitative PCR reaction that allows relative quantification of the initial protein concentration[105]. Olink panels include internal and external controls on each plate: two controls of the immunoassay (two non-human proteins), one control of oligonucleotide extension (an antibody linked to two matched oligonucleotides for immediate proximity, independent of antigen binding) and one control of hybridized oligonucleotide detection (a pre-made synthetic double stranded template), as well as an external, between-plate, control (http://www.olink.com/; accessed: 19th June 2016).

Prior to analysis, we excluded proteins with fewer than 200 samples with measurements above the limit of detection of the assay. Of the 268 unique proteins reported by Olink, 253 passed this threshold in ORCADES, and 252 in CROATIA-Vis, with an intersect of 251 proteins. Protein values were inverse-normal rank-transformed prior to subsequent analysis.

The subunits of IL27 are not distinguished in Olink's annotation (Q14213, *EBI3*; and Q8NEV9, *IL27*). However, it has only one significant locus, local to the *EBI3* gene (lead variant, rs60160662, is within 16kb). Therefore, *EBI3* (Q14213) was selected as representative for this protein when discussing pQTL location (local/distal) so as to avoid double counting.

The CVD2, CVD3, and INF panels are commercially available from Olink. The proteins on these panels were selected by Olink due to *a priori* evidence of involvement in cardiovascular and inflammatory processes. Two proteins, CCL20 and BDNF, have been removed at the request of Olink (due to issues with the assay).

Detection of pQTL. Genome-wide association of these proteins was performed using autosomes only. Analyses were performed in three-stages. (1) a linear regression model was used to account for participant age, sex, genotyping array (ORCADES only), proteomics plate, proteomics plate row, proteomics plate column, length of sample storage, season of venepuncture (ORCADES only), and the first 10 principal components of the genomic relationship matrix. Genotyping array and season of venepuncture are invariant in CROATIA-Vis and therefore were not included in the model. (2) Residuals from this model were

corrected for relatedness, using GenABEL's[100] polygenic function and the genomic relationship matrix, to produce GRAMMAR+ residuals. Outlying GRAMMAR+ residuals (absolute z-score >4) were removed and the remainder rank-based inverse-normal transformed. (3) Genome-wide association testing was performed using REGSCAN v0.5[106].

Genome-wide association results were clumped by linkage disequilibrium using PLINK v1.90[101,102]. Biallelic variants within ±5Mb and r^2 >0.2 to the lead variant (smallest p-value at the locus) were clumped together, and the lead variant is presented. r^2 was derived from all European populations in 1,000 Genomes[92,93].

We have chosen to describe pQTL as *local-* or *distant-* so as to distinguish naming based on genomic location from that based on mode of action i.e. *cis-* (acting on the same DNA molecule) and *trans-* (acting via some diffusible mediator). That is, most *local-* variation may well act in *cis* but not necessarily so.

Mendelian Randomisation. In the context of proteome-by-phenome MR, a DNA variant (a single nucleotide polymorphism in this case) that influences plasma protein level is described as an 'instrumental variable', the protein as the 'exposure variable', and the outcome phenotype as the 'outcome variable'.

The lead-SNP with the lowest p-value meeting the following criteria was used as the instrumental variable for each protein:

(1) Minor allele frequency >1% in both ORCADES and CROATIA-Vis cohorts.

(2) An imputation info score (SNPTEST v2) of >0.95 in both ORCADES and CROATIA-Vis.

(3) Located within ±150kb of the gene coding for the protein (start and end coordinates of the gene as defined by Ensembl GRCh37[107]).

(4) Significant (as defined below) SNP:protein link in both the discovery and replication cohorts.

Lead-SNP selection was performed using the discovery (CROATIA-Vis; p-value <$5x10^{-8}$) cohort; replication was defined based on a Bonferroni correction for the number of significant lead-SNPs present in the discovery cohort (CROATIA-Vis). In order to avoid a 'winner's curse', genome-wide association effect size estimates and standard errors from the replication cohort (ORCADES) were used for MR.

We perform MR as a ratio of expectations, using up to second-order partial derivatives of the Taylor series expansion for effect size estimates, and up to first-order for standard errors (Delta method)[108]:

(1) $$\beta_{YX} \approx \frac{\beta_{YZ}}{\beta_{XZ}} \left(1 + \frac{se^2_{XZ}}{\beta^2_{XZ}}\right)$$

(2) $$se_{YX} \approx \sqrt{\frac{se^2_{YZ}}{\beta^2_{XZ}} + \frac{\beta^2_{YZ} \cdot se^2_{XZ}}{\beta^4_{XZ}}}$$

(3) $$p_{YX} = 2\Phi\left(-|\beta_{YX}|/se_{YX}\right)$$

where β_{ij} is the causal effect of *j* on *i*, se_{ij} is the standard error of the causal effect estimate of *j* on *i*; subscript *X* is the exposure, *Y* the outcome trait, and *Z* the instrumental variable. Φ is

the cumulative density function of the standard normal distribution. This method is identical to that of SMR[27] apart from the second term in the bracket of Equation 1 (resulting from the inclusion of second-order partial derivatives). An FDR of <0.05 was considered to be significant. FDR estimations were performed separately on those results derived from GeneAtlas and those derived from studies in Phenoscanner.

DNA variant to trait association: GeneAtlas. UK Biobank has captured a wealth of information on a large – approximately 500,000 individuals – population cohort that includes anthropometry, haematological traits, and disease outcomes. All 778 outcome traits from UK Biobank in GeneAtlas (http://geneatlas.roslin.ed.ac.uk/; Canela-Xandri et al. (2018)[10]) were included. The analysis method of all 778 traits was as described for 717 in Canela-Xandri et al. (2017)[29]. For each protein, the lead (lowest DNA variant-protein association p-value in the discovery cohort) biallelic (Phase 3, 1,000 Genomes[92,93]) variant meeting the criteria above and an imputation info score >0.95 in UK Biobank, was selected for each protein, and MR performed.

DNA variant to trait association: Phenoscanner. Phenoscanner[67,68] was used to highlight existing GWA studies for inclusion. For each protein, the lead (lowest DNA variant-protein association p-value in the discovery cohort) biallelic (1,000 Genomes[92,93]) meeting the criteria above was selected. rs545634 was not found in the Phenoscanner database and was therefore replaced with the second most significant variant meeting the above criteria: chr1:15849003. Phenoscanner was run with the following options: Catalogue: 'Diseases & Traits', p-value cut-off: '1', Proxies: 'None', Build '37'. The results from those studies that

returned a value for all input variants were kept and MR performed. Phenoscanner (http://www.phenoscanner.medschl.cam.ac.uk/information/; accessed 25 Sep 2018) state that they report all SNPs on the positive strand. Given this, alleles were harmonised as required. No attempt to harmonise based on allele frequency was made; therefore, the direction of effect of C/G and A/T SNPs should be interpreted with care. Results from 20 additional studies were obtained, corresponding to 68 outcomes.

HEIDI. Heterogeneity in dependent instruments (HEIDI) analysis[27], is a method of testing whether the MR estimates obtained using variants in linkage disequilibrium with the lead variant are consistent with a single causal variant at a given locus (Figure 1d). HEIDI analysis was performed using software provided at https://cnsgenomics.com/software/smr/ (accessed 28 Aug 2018; v0.710). We used pQTL data from ORCADES for assessment as the exposure. Biallelic variants from the 1,000 Genomes[92,93] (European populations: CEU, FIN, GBR, IBS, and TSI) were used as the linkage disequilibrium reference. We used the default 'cis-window' of 2000kb, and a maximum number of variants of 20 (as is the default value for the software).

We performed HEIDI analysis of all exposure-outcome links that were found to be significant (FDR <0.05) using outcomes from GeneAtlas (n =271), as well as links found to be MR significant (FDR <0.05) with CAD from the meta-analysis of van der Harst[61], and for SHPS1 and schizophrenia[51].

We applied the following filters for variants to be included in the analysis: minor allele frequency MAF >0.01 and, in the GeneAtlas and ORCADES data, an imputation info score of >0.95.

Comparison to eQTL

Result for all SNP:gene pairs analysed in whole blood were downloaded from GTEx[13] (v7). Results were extracted for the instrumental variables and the genes encoding their proteins for the 64 proteins for which an instrumental variable was successfully identified in this study. Matching was based on Ensembl Gene ID, and variant chromosome, position, and alleles (GRCh37).

Comparison to plasma pQTL using an orthogonal, aptamer-based, method

The supplementary data files for Sun et al[14] were downloaded on 04 Sep 2019. pQTL identified were extracted for the 64 proteins for which an instrumental variable was successfully identified in this study. Proteins were matched based on an exact UniProtID match. The LD (r^2) between the lead locally-acting (as defined above) and 'cis-acting' (as defined by Sun et al.) SNP identified for each protein was calculated using the European populations from the 1,000 Genomes project (as described above) using PLINK v1.90[101,102].

Links to existing drug therapies

Protein names were matched to ChEMBL IDs using the UniProtID mapping API (https://www.uniprot.org/help/api_idmapping; accessed 27 Oct 2019). ChEMBL[109] was searched programmatically using the ChEMBL web resource client in Python 3.6 (https://github.com/chembl/chembl_webresource_client; accessed 27 Oct 2019).

Declarations

Ethics approval and consent to participate:

ORCADES: The study was approved by Research Ethics Committees in Orkney and Aberdeen (North of Scotland REC, 26/11/2003).

CROATIA-Vis: The study received approval from the relevant ethics committees in Scotland (South East Scotland Research Ethics Committee, REC reference: 11/AL/0222) and Croatia (University of Split School of Medicine Ethics committee, Class:003-08/11-03/-005 No.: 2181-198-03-04/10-11-0008).

All participants gave written informed consent and both studies complied with the tenets of the Declaration of Helsinki.

Consent for publication: Not applicable, no individual level data presented.

Availability of data and material:

Datasets supporting the conclusions of this article are included within the article and its additional files. In addition, summary level data will be made available on publication. However, there is neither research ethics committee approval, nor consent from individual

participants, to permit open release of the individual level research data underlying this study. Please contact the QTL Data Access Committee (accessQTL@ed.ac.uk) for further information if required.

Competing interests: The authors declare that they have no competing interests.

Funding:

- ADB would like to acknowledge funding from the Wellcome PhD training fellowship for clinicians (204979/Z/16/Z), the Edinburgh Clinical Academic Track (ECAT) programme.
- TB, YZ, CA, PN, JFW, VV, CHay, CPP and CHal are supported by MRC University Unit Programme Grants to the Human Genetics Unit (MC_PC_U127592696, MC_UU_12008/1, MC_UU_00007/10 and MC_UU_00007/15)
- AT, OC-X and KR acknowledge funding from the MRC (MR/R025851/1, MR/N003179/1).
- CHal, JKB, AT, and KR acknowledge funding from BBSRC Institute Strategic Programme grants to the Roslin Institute (BBS/E/D/30002275, BBS/E/D/30002276, BBS/E/D/10002071, BBS/E/D/20002172, BBS/E/D/20002174).
- PKJ would like to acknowledge funding from the Axa research fund.
- JKB acknowledges funding support from a Wellcome-Beit Prize Intermediate Clinical Fellowship (103258/Z/13/Z,A), and the UK Intensive Care Foundation.
- The Orkney Complex Disease Study (ORCADES) was supported by the Chief Scientist Office of the Scottish Government (CZB/4/276, CZB/4/710), a Royal Society URF to JFW, the MRC Human Genetics Unit quinquennial programme "QTL in Health and Disease", Arthritis Research UK and the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947).

- The CROATIA-Vis study was funded by grants from the Medical Research Council (UK) and Republic of Croatia Ministry of Science, Education and Sports research grants (108-1080315-0302).

Authors' contributions (alphabetical order):

Computational analysis and/or critical interpretation of the results: ADB, AT, CA, CHal, CPP, DWC, OC-X, PN, JKB, KR, PKJ, TSB, YZ.

Conceived, designed and/or managed the cohort data: AFW, CHal, CHay, HC, IR, JFW, VV.

Wrote the manuscript: ADB, CHal, CPP, JKB.

All authors have read and approved the final manuscript.

Acknowledgements:

- A debt of gratitude is owed to all the participants in all cohorts used, without whom this work would not have been possible.
- This research has been conducted using the UK Biobank Resource under project 788.
- The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal on 04 Sep 2019.
- We would like to acknowledge the invaluable contributions of the research nurses in Orkney, the administrative team in Edinburgh and the people of Orkney. DNA extractions were performed at the Wellcome Trust Clinical Research Facility in Edinburgh.

- We would like to acknowledge the staff of several institutions in Croatia that supported the field work, including but not limited to The University of Split and Zagreb Medical Schools, the Institute for Anthropological Research in Zagreb and Croatian Institute for Public Health. Genotyping was performed in the Genetics Core of the Clinical Research Facility, University of Edinburgh.

Page intentionally blank.

Figure 1. Proteome-by-phenome Mendelian Randomisation.

A) Genome-wide associations of the plasma concentrations of 249 proteins from two independent European cohorts (discovery and replication) were calculated. The plot shows pQTL position against chromosomal location of the gene that encodes the protein under study for all replicated pQTLs. The area of a filled circle is proportional to its -log10(p-value) in the replication cohort. Blue circles indicate pQTLs ±150kb of the gene ('local-pQTLs'); red circles indicate pQTLs more than 150kb from the gene. B, C) Local-pQTLs of 64 proteins were taken forward for proteome-by-phenome MR analysis. These were assessed against 778 outcome phenotypes from GeneAtlas[29] (panel B; UK Biobank) and 68 phenotypes identified using Phenoscanner[67,68] (panel C). In each set of results an FDR of <0.05 was considered significant. D) Heterogeneity in dependent instruments (HEIDI[27]) testing was undertaken for MR significant results from GeneAtlas (n =271). This test seeks to distinguish a single causal variant at a locus effecting both exposure and outcome directly (as in i) or in a causal chain (as in ii), from two causal variants in linkage disequilibrium (as in iii), one affecting the exposure and the other effecting the outcome.

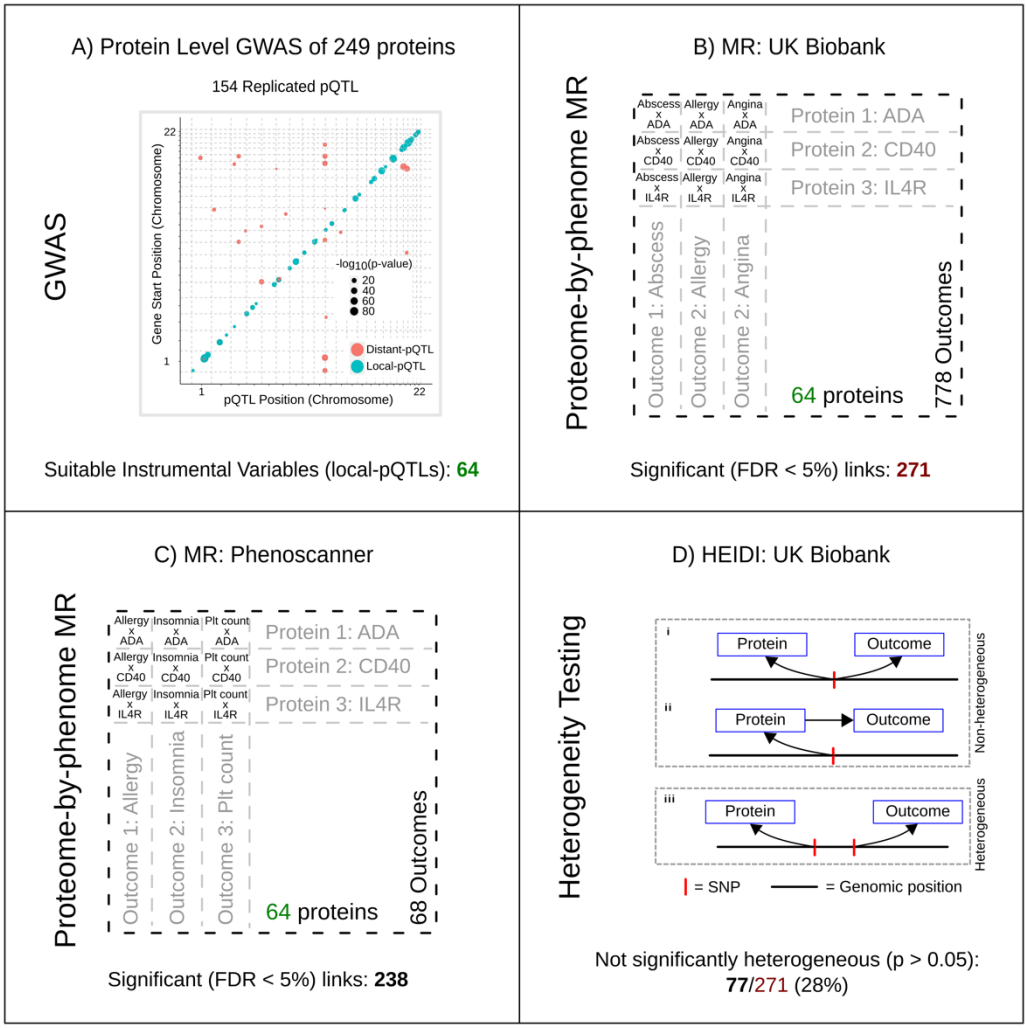

A) Protein Level GWAS of 249 proteins
GWAS
154 Replicated pQTL
Gene Start Position (Chromosome)
pQTL Position (Chromosome)
-log10(p-value)
Distant-pQTL
Local-pQTL
Suitable Instrumental Variables (local-pQTLs): 64
B) MR: UK Biobank
Proteome-by-phenome MR
Protein 1: ADA
Protein 2: CD40
Protein 3: IL4R
Outcome 1: Abscess
Outcome 2: Allergy
Outcome 2: Angina
64 proteins
778 Outcomes
Significant (FDR < 5%) links: 271
C) MR: Phenoscanner
Proteome-by-phenome MR
Protein 1: ADA
Protein 2: CD40
Protein 3: IL4R
Outcome 1: Allergy
Outcome 2: Insomnia
Outcome 3: Plt count
64 proteins
68 Outcomes
Significant (FDR < 5%) links: 238
D) HEIDI: UK Biobank
Heterogeneity Testing
Protein
Outcome
Non-heterogeneous
Heterogeneous
= SNP
= Genomic position
Not significantly heterogeneous (p > 0.05):
77/271 (28%)

Figure 2. Significant (FDR <0.05) proteome-by-phenome MR protein-outcome causal inferences: disease subset.

MR significant (FDR<5%) protein-disease outcome results.

a) All MR significant (FDR<5%) protein-disease outcome results for outcomes from the Phenoscanner[67,68] studies (see key for details).

b) All MR significant (FDR<5%) protein-disease outcome results for outcomes from GeneAtlas[29]. An asterisk indicates MR estimates that are *not* significantly heterogeneous upon HEIDI testing (see key for details).

c) Key. From the outside in: HGNC symbol of the protein (exposure); disease outcome; key colour (matching the protein name in the outer ring); bar chart of the signed squared beta estimate divided by the squared standard error of the MR estimate, using pQTL data from the discovery cohort (CROATIA-Vis); bar chart of the signed squared beta estimate divided by the squared standard error of the MR estimate, using pQTL data from the replication cohort (ORCADES). Central links join identical outcomes for which more than one protein was found to be MR significant. The colour of the links indicates similar outcome groups, e.g. thyroid disease.

The key to the outcome descriptions is detailed further in Supplementary Table 8 and Supplementary Table 9.

d) Example concordance (due to sample overlap) plot for all proteins with significant MR evidence in GeneAtlas for causal roles in asthma (IL1RL1, IL1RL2, IL2RA, IL4R, IL6R). GeneAtlas traits are on the left. Phenoscanner traits are on the right. Thickness of connecting lines is proportional to -$\log_{10}$(p-value). The Phenoscanner studies included here are derived from [47,49,50,53,61,64–66], of which [49,61,65,66] include at least some part of the UKBB

data. However, [49,65,66] use only data from the first phase (~150,000 individuals) genotype release from UK Biobank..

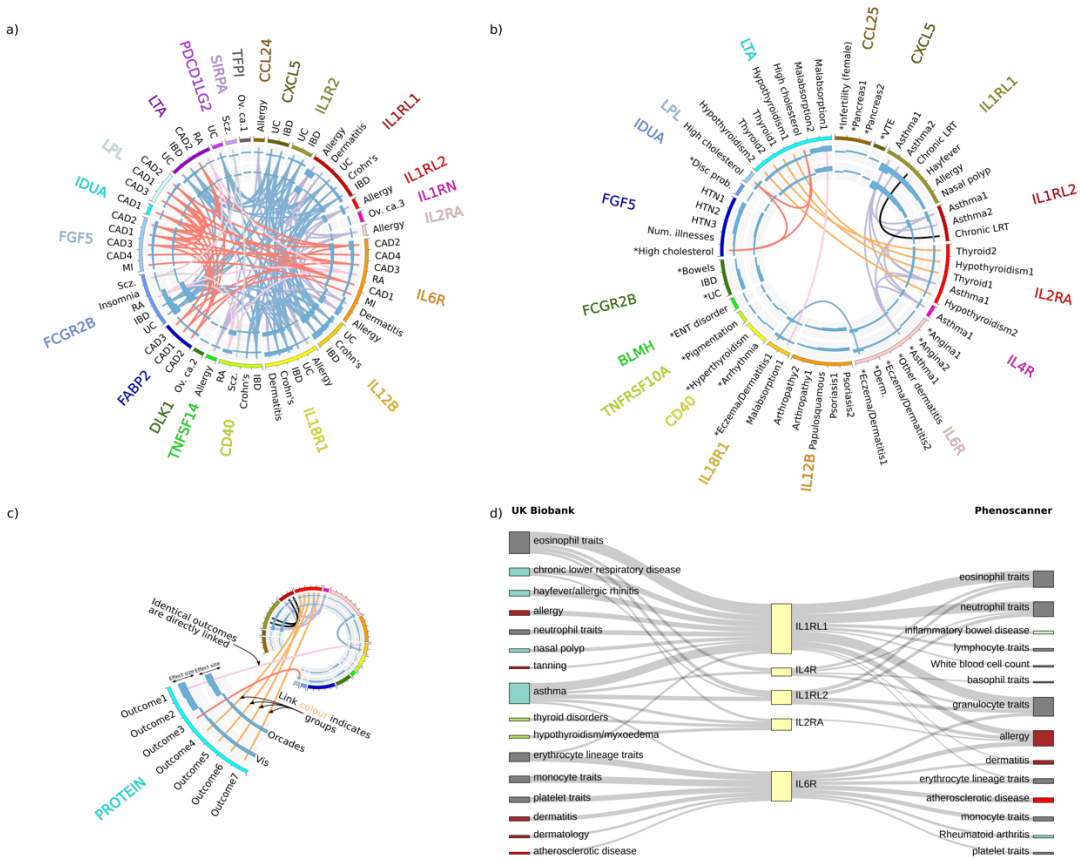
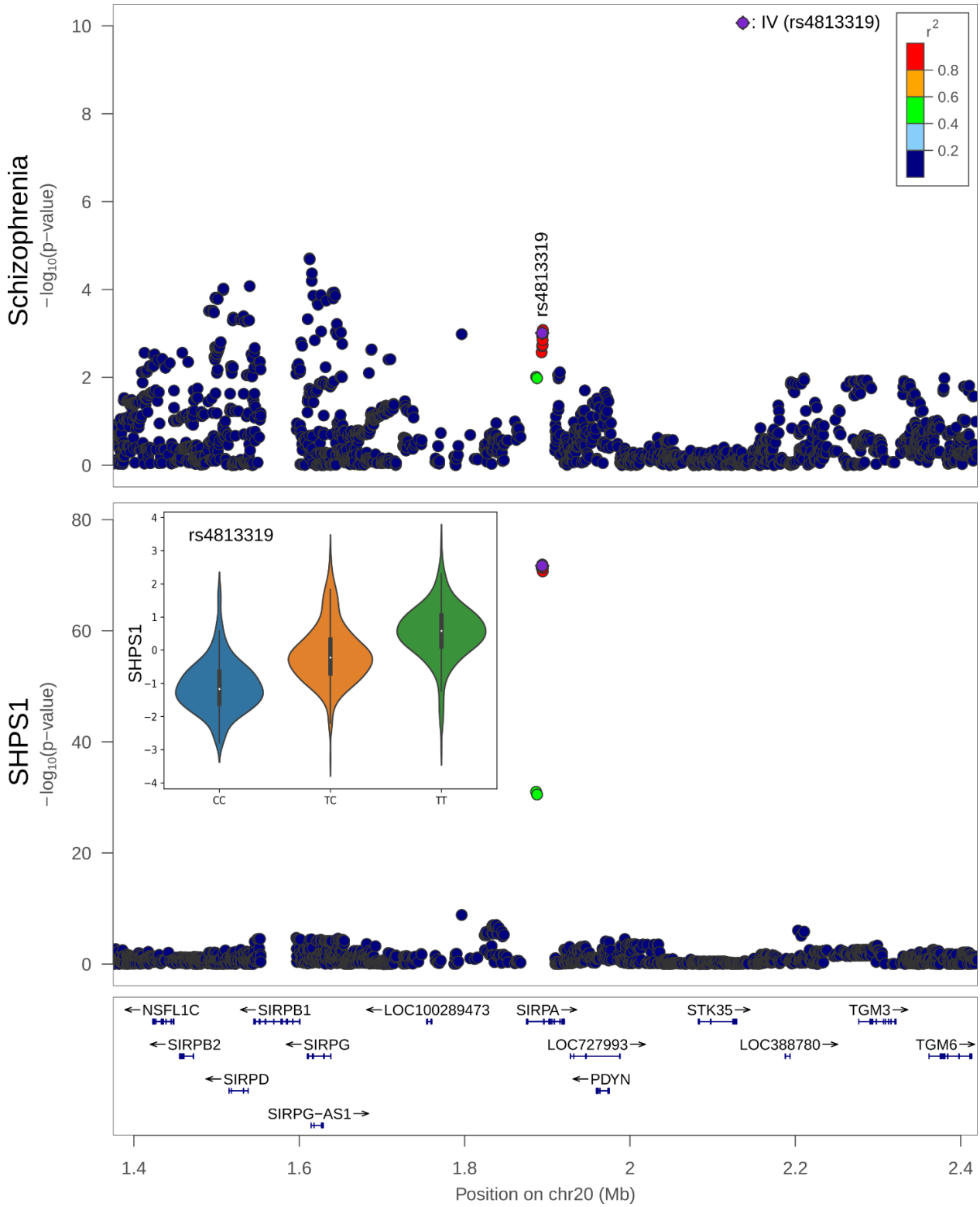


Figure 3: Co-localisation of SHPS1 (encoded by SHPS1: synonym SIRPA) and schizophrenia DNA associations.

Upper panel, locuszoom[110] of the region surrounding *SHPS1* and the associations with schizophrenia[51]; lower panel, associations with SHPS1. Lower panel inset, the relative concentration of SHPS1 across the 3 genotypes of rs4813319 – the DNA variant used as the instrumental variable (IV) in the MR analysis: CC, CT, and TT.


Schizophrenia
−log10(p−value)
: IV (rs4813319)
r2
0.8
0.6
0.4
0.2
rs4813319
SHPS1
−log10(p−value)
rs4813319
SHPS1
CC
TC
TT
NSFL1C
SIRPB1
LOC100289473
SIRPA
STK35
TGM3
SIRPB2
SIRPG
LOC727993
LOC388780
TGM6
SIRPD
PDYN
SIRPG−AS1
Position on chr20 (Mb)

3. Cellular mass-spectrometry based proteomics

Introduction

Peripheral blood mononuclear cells (PBMCs) are a site of active gene expression in blood [111]. Recent endeavours have begun to unravel the depths of the human proteome and have found many genetic variants influencing protein abundance. If one is happy to accept that protein abundance cannot influence genotype of an individual, then these are causal links. However, due to the relative ease of sampling, previous studies have generally focused on plasma proteomes [14,41–43], or model cellular systems (such as induced pluripotent stem cells [19]). Unfortunately, plasma protein abundance derives from multiple sources and is unlikely to be representative of any single cell-type or tissue, and it remains an open question as to how representative the plasma proteome is of any primary tissue.

PBMCs, composed of lymphocytes and other mononuclear cells (natural killer cells, dendritic cells, and monocytes), play a central role in the human immune-system. They are instrumental in both the pathogenesis and prevention of many important diseases. They are deeply entwined in the genesis of auto-immune disease, and are directly involved in the host-response to infection. In addition, they are also the target tissue of many globally important diseases, especially viral infections, including human immuno-deficiency virus (HIV), Epstein-Barr virus (EBV), and human T-cell lymphotropic virus (HTLV). PBMCs therefore represent an important physiologically relevant tissue – the proteome of which I have assessed directly, and at scale.

The difficulty with traditional measurements of protein abundance, for example western blotting, is that they are not easily scalable. Current technologies used to assess the proteome at population scales can broadly be categorised into two groups: antibody-based assays, and aptamer-based assays. As the name implied, antibody-based assays, such as the Olink platform (as used in Chapter 2), use one or more antibodies targeted at the proteins of interest. In the case of the Olink platforms, the array has a matched pair of antibodies for each protein, each linked to one of a pair of oligonucleotides. Binding of both the antibodies to the protein brings the paired oligonucleotides into close proximity. This permits hybridization, and these oligonucleotides then form the basis of a quantitative PCR reaction. This then allows relative quantification (between samples) of protein abundance [105].

As with the Olink assays, aptamer-based assays aim not to measure protein abundance directly, but to measure DNA abundance as a proxy. 'Slow off-rate modified aptamer' (SOMAmer) technology [112] involves oligonucleotide aptamers with protein-like side-chains that binds to a target molecule (in this context a protein). These are available commercially as part of an aptamer-based multiplex protein assay (SOMAscan) and has been used to form the basis of the largest analysis of the human plasma proteome to date [14]. However, whilst these technologies are useful, they are not a panacea. Specificity to their putative targets, especially in complex mixtures, is difficult to guarantee. Interestingly, in order to assess the binding specificity, academia and industry often recourse to mass-spectrometry based methods [113].

In this chapter I present, a direct assessment of the cellular proteome of PBMCs with high-performance liquid chromatograph mass-spectrometry (HPLC-MS/MS). This enables a

deeper assessment of the proteome than is possible with current antibody-, or aptamer-based technologies, and is not affected by the difficulties of binding-specificity.

Here I present the first phase of a much larger study, demonstrating the feasibility of assessing the cellular proteome of PBMCs, using HPLC-MS/MS, at a population scale. I describe the results of the first 251 samples processed. Even at this scale, this is, to the best of my knowledge, the first study to systematically assess paired genomic and proteomic information to this depth, at a population-scale, in primary human cells.

Currently the results include assessment of the genome-wide association of 4,433 proteins, each measured in 100 or more samples of PBMCs from 251 individuals. In addition, the number of proteins, and the proportion of samples within which they are detected, is expected to improve as the project progresses.

In this chapter, I perform GWA of the 4,433 proteins and assess concordance with eQTL and plasma pQTL, using results from GTEx [13] and Sun et al. [14], respectively. Finally, rheumatoid arthritis was selected as a representative autoimmune disease – a group of diseases for which PBMCs are likely a relevant set of cell-types – and two-sample Mendelian Randomisation performed using the results of Okada et al. [53].

Methods

The PBMC samples for the main part of this study were obtained from Generation Scotland (GS application number for this project: GS18318). Generation Scotland: Scottish Family Health Study (GS) is a population- and family-based cohort from the Scottish population. Participants were recruited between 2006 and 2011 and blood samples obtained at the time of recruitment. The cohort has previously been described in detail elsewhere [114]. Participants provided written informed consent. Ethical approval was provided by the East of Scotland Research Ethics Service committee on research ethics (REC references 15/ES/0040).

Peripheral-blood mononuclear cell preparation

PBMC samples were available for 862 genotyped individuals from Generation Scotland. PBMCs were separated from approximately 5ml blood in acid-citrate-dextrose using density gradient separation (Histopaque-1077; Sigma-Aldrich) by the European Collection of Authenticated Cell Cultures (ECACC) using a standardised protocol. PBMCs were frozen in a rate-controlled manner and stored in liquid nitrogen (in foetal-calf serum with 10% DMSO) until withdrawn for this study. In addition, further PBMC samples were obtained for protocol optimisation, using an analogous extraction technique to that used in Generation Scotland, at the Roslin Institute, ethical approval for which was granted by South East Scotland Research Ethics Committee 02 (REC reference 11/AL/0168).

Genotyping

Genotyping and downstream processing of genotypes had already been performed prior to the commencement of this study. In brief, it was performed at the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland using the HumanOmniExpressExome-8 v1.0 or v1.2 BeadChips with Infinium chemistry (Illumina). Genotypes were processed using the GenomeStudio Analysis software v2011.1 (Illumina) and called using Beadstudio-Gencall v3.0 (Illumina). The details of blood collection and DNA extraction are provided elsewhere [115]. Quality control: individuals with <98% call rate, SNPs with <98% call rate, or a Hardy-Weinberg equilibrium p-value $<1x10^{-6}$. The resulting data set was merged with data from the 1000 Genomes project [116] and a principal component analysis performed using GCTA [117]. Individuals more than six standard deviations away from the mean of principal component 1 and principal component 2 were removed. SNPs with a minor allele frequency <0.05 were removed. Following this, 519,798 genotyped autosomal SNPs remained.

Sample preparation for mass-spectrometry

The MRC Institute of Genetics and Molecular Medicine (IGMM) mass-spectrometry team already had protocols for sample preparation for mass-spectrometry for small numbers of samples. In collaboration with them, I undertook optimisation of their existent protocols to ensure the practicality of performing this in a standardised way, at scale. Initial data were obtained using the samples from the Roslin Institute, and subsequently using ungenotyped samples from Generation Scotland. Protocols were followed, as closely as was practical, with the following notable exceptions: 1) The pooled standard and Batch 1 were originally

sonicated with a BioRuptor for 20 min. However, during the processing of Batch 1, this was found to be ineffective at fully disrupting the nucleic acid. Each sample from Batch 1 was therefore further sonicated using a Soniprep 150 for 10sec at 5μm amplitude. As reflected in the protocol, the Soniprep 150 was then used as the principal method for subsequent batches; 2) Batch 4 received an additional wash with 10ml phosphate buffered saline per sample; and 3) The digest of batches 1-3, and 6 were repeated using frozen cell lysate rather than directly from cell lysate on dry ice. The final protocols are found in the supplementary materials, and are summarised below.

Cell preparation and lysis

Frozen cell pellets were retrieved from liquid nitrogen and washed three times in chilled phosphate buffered saline (once with 9ml, twice with 10ml) prior to lysis in 40µl of 6M guanidine hydrochloride with 100mM tris(hydroxymethyl)aminomethane (tris) – 'lysis buffer'. 2µl of sample was taken for a protein assay (Pierce BCA protein assay, Thermo-Fisher; Catalog number 23227). Based on the results of this assay, protein concentrations were standardised to a ceiling of 15µg of protein per well of a 96-well PCR plate in 20µl of lysis buffer. Following lysis, samples were sonicated to ensure disruption of nucleic acid. In addition to the samples of the batch, per-plate, an additional within-batch repeat and up to two between-batch repeats were included. Samples were prepared in batches: 19 batches of 43, and one of 45.

Protein digest and peptide preparation

Each sample was reduced with tris-carboxyethylphosphine (TCEP; 1µl 100mM), alkylated with chloroacetamide (CAA; 1µl 200mM), and heated for 5 minutes to 90-95°C. After cooling, samples were diluted two-fold with 20µl 100mM tris (pH 8.5) and digested overnight with mass-spectrometry grade lysyl endopeptidase (300µg per well. Wako: reference 121-05063; lot numbers CAR3124 and CAR3125) at 37°C. Samples were then diluted a further three-fold with 80µl 50mM tris (pH 8.5) and digested for 4 hours with trypsin (150µg per well; Pierce (Thermo Fisher) reference 90058, lot TG269839), again at 37°C. Digestion was stopped by the addition of 16µl 10% trifluoroacetic acid (TFA) per well. Peptides were de-salted on C18 columns [118,119]. Columns were activated with 15µl methanol, washed with 50µl 0.1% TFA (pre- and post- sample loading), and eluted with 40µl 80% acetonitrile (ACN) + 0.1% TFA. Following elution, samples were dried and resuspended in 14µl mass-spectrometry grade water. 5µl of each sample was taken for a peptide assay (Pierce Quantitative Colorimetric Peptide Assay (Thermo Fisher) Catalog number 23275). The remaining 9µl of sample were acidified with 1µl 1% TFA and stored frozen prior to mass-spectrometry analysis.

Pooled standard and library generation.

A pool was generated from 20 individuals from GS. The pool was aliquoted into 20µl aliquots (containing 15µg protein per aliquot) and prepared as the primary samples. However, following resuspension in water, the samples were again pooled. The pool was then split: part for use as a pooled standard for running with each batch, and part for the

genesis of a peptide library. 20µl of the pooled standard was run with each batch. Note that the peptide library was not included in the results of the first phase analysis presented here.

Mass-spectrometry machine protocol

LC-MS/MS was performed on a Thermo Ultimate 3000 RSLC Nano UPLC coupled to a Thermo Fisher Q Exactive plus mass-spectrometer (Thermo Fisher). Samples were directly injected from a 96-well plate onto an Aurora UHPLC column from IonOpticks (Ion Opticks Pty Ltd). A Proxeon nano-spray ionisation source (Proxeon Biosystems) with a capillary temperature of 250°C and an optimised voltage of 1.4-1.7kV was used. A 120 minute gradient (2%-30% B in 110 min, 30-45%B in the next 10 min; A=2% acetonitrile, B=80% acetonitrile, 0.5% acetic acid throughout; the composition was raised to 100% B in 7 minutes after the analytical gradient to wash the column, and total equilibration time was 20 minutes), data-dependent acquisition, was run with a scan range of 350 to 1400 m/z using the Orbitrap at a resolution of 70,000 in profile mode. The top 24 parent peaks were selected for fragmentation. HCD fragmentation was performed with a normalised collision energy of 26 and spectra were acquired in centroid mode at a resolution of 17,500. Charge-states accepted for MS2 were 2-5, peptide match was preferred and dynamic exclusion was 30 seconds.

Data processing

Data search and annotation.

Data were processed using MaxQuant [33,34] (v1.6.5.0), matching against UniProt human (9606) reference proteome (5640) release 2019_01 [120] The search was performed for trypsin-digested peptides with up to two permitted missed cleavages, the fixed modification carbamidomethylation (C), and the variable modifications of oxidation (M) and acetylation (Protein N-terminus). MaxQuant match-between-runs and label-free quantification were used [121]. The false discovery rate of the peptide-spectrum matches was calculated using a decoy-target approach. A threshold of 0.01 used for the search. Match between run parameters: matching time window: 0.7 minutes; alignment time window: 20 minutes.

Data transformation

Label-free quantification (LFQ) values were quantile-normal transformed per sample. Missing data were imputed into the lower tail of the normal distribution. Proteins that were measured in fewer than 100 samples were not included in further analysis.

Genome-wide association

A linear mixed model was fit using GEMMA [122], for each SNP the phenotype was fit against a model containing the SNP, an intercept, other fixed effects (age, age^2, sex, batch, and haemoglobin subunit alpha abundance), and a standardized whole-genome genomic relationship matrix as a random effect.

An initial genome-wide screen was performed as well as a subsequent local-pQTL analysis limited to ±1Mb window surrounding the gene start and end positions (as defined by

Ensembl GRCh37 [107]). I considered genotyped SNPs with a minor allele frequency >0.05 in the set of 251 individuals. A false discovery rate of <0.05 was considered to be significant in the *locally*- acting set.

Mapping to GTEx (v6p) and plasma pQTL

All per gene significant (as defined by GTEx) *locally*- acting (±1Mb from the transcription start site) eQTL in GTEx (v6p) [13] were searched for the lead-SNP of each significant (FDR <0.05) *locally*- acting (as defined above) pQTL identified in this chapter. Matches were based on chromosome, position, and alleles and ENSG ID.

The supplementary tables of Sun et al. [14] were downloaded from the publisher's website. Proteins were matched between the PBMC proteomics data and Sun et al. based on an exact UniProt ID match.

Rheumatoid Arthritis Mendelian randomisation

The full summary statistics of Okada et al. [53] (trans-ethnic) were searched for all the lead-SNP of each significant (FDR <0.05) *locally*- acting (as defined above) pQTL identified in this chapter. Matches were based on rsID. Attempted replication of those SNPs significantly associated (Bonferroni correction) with rheumatoid arthritis (RA) in Okada et al. was attempted in the GeneAtlas [10], and consistency of causal effect of protein on disease assessed using MR (Equation 3.1 and Equation 3.2).

Equation 3.1: Mendelian Randomisation β estimate, Delta method (first order expansion)

$$\beta_{MR} = \frac{\beta_{SNP:RA}}{\beta_{SNP:PROT}}$$

Equation 3.2: Mendelian Randomisation standard error estimate, Delta method (first order expansion)

$$\sigma_{MR} = \sqrt{\frac{\sigma^2_{SNP:RA}}{\beta^2_{SNP:PROT}} + \frac{\beta^2_{SNP:RA}\sigma^2_{SNP:PROT}}{\beta^4_{SNP:PROT}}}$$

$\beta_{SNP:RA}$ and $\beta_{SNP:PROT}$ are the effect estimates of the SNP on rheumatoid arthritis, and protein abundance respectively.

$\sigma_{SNP:RA}$ and $\sigma_{SNP:PROT}$ are the standard errors of the effect estimate of the SNP on rheumatoid arthritis and the protein abundance respectively.

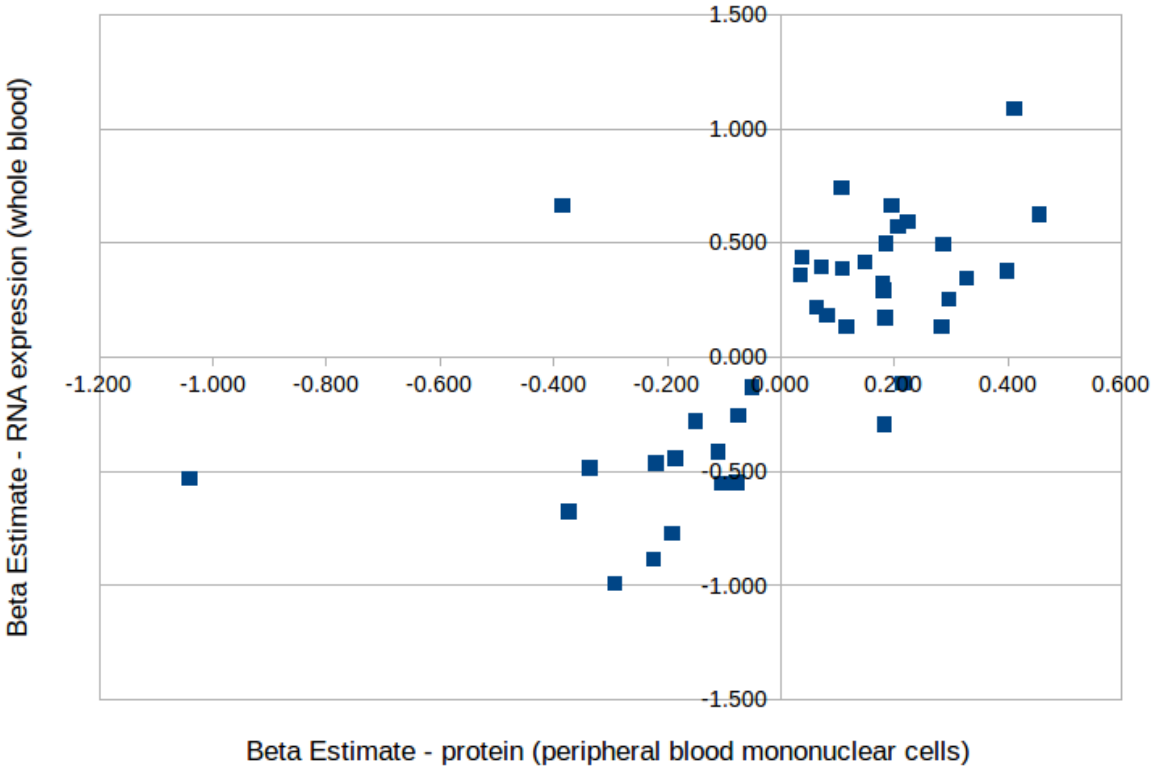
Results

I identified 4,433 proteins in PBMCs, each in 100 or more samples, at a peptide-spectrum match false-discovery rate of <0.01. At a genome-wide significance threshold (Bonferroni threshold; p-value < $5x10^{-8}/4,433$), I identified 25 proteins with one or more pQTL. Of these, 21 mapped to an autosome (Ensembl GRCh37 [107]) and, in all 21 instances, were located within 1Mb of a gene encoding the protein. When focussed on *local* variation (±1Mb of the mapped gene(s) encoding the protein under study), I identify 108 proteins with one or more significant (false discovery rate (FDR) <0.05) pQTL. These 108 proteins map to 113 Ensembl

Gene IDs (GRCh37) [107]. Of these, the SNP:ENSG ID pair was also called a significant per ENSG ID expression quantitative trait locus (eQTL) in whole blood from GTEx (v6p) [13] in 38.1% (43/113; Supplementary Table 10), with a consistent direction of effect in 92.7% (38/41; Figure 3.1; C/G and T/A SNPs excluded from directional concordance analysis). The co-efficient of determination of the linear trend line was 0.45, broadly similar with estimates from previous studies [19].

Figure 3.1: Comparison of peripheral blood mononuclear cell pQTL and whole blood eQTL.

Beta estimates from peripheral blood mononuclear cells pQTL (FDR <0.05, *locally*- acting, lead-SNPs; unitless) that were also whole blood eQTL (GTEx v6p [13]; unitless). N=41.



Rheumatoid arthritis

Rheumatoid arthritis is an inflammatory arthropathy affecting approximately 1% of the worlds' population. It is not only a destructive joint disease, but can also have serious systemic effects, for example interstitial lung disease. Given the inflammatory nature of this globally important condition, it was felt to be a good test case for MR using immune-cell specific protein pQTLs.

The 108 significant (FDR <0.05) *locally*- acting PBMC pQTL lead-SNPs were looked up in the results of a large rheumatoid arthritis (RA) genome-wide association (GWA) study [53]. Of these, 104 were reported in the outcome study, and 7 SNPs were found to be significant (Bonferroni correction, p-value <0.05/104), corresponding to the pQTL of HLA-DQA1, NELFE, PADI4, HLA-B, RNASET2, PADI2, and HLA-F. Of these, two of the SNP:RA associations (the SNPs tagging PADI2 and HLA-F) were not genome-wide significant (p-value >$5x10^{-8}$) in the RA GWA results.

The SNPs tagging HLA-DQA1, NELFE, and PADI4 all replicated (Bonferroni correction, 0.05/7; associated with rheumatoid arthritis (ICD10: M05 Seropositive rheumatoid arthritis) and a consistent direction of effect) in GeneAtlas [10] (p-values: $6.65x10^{-12}$, $1.15x10^{-8}$, and $3.23x10^{-3}$, respectively). Interestingly PADI2 and PADI4 are both (and the only) members of the 'histone H3-R26 citrullination' pathway (GO:0036413). Citrullination of proteins being a key feature of the pathogenesis of RA, indeed, anti-cyclic citrullinated peptide (CCP) antibodies are used in clinical practice as part of the diagnostic work-up for RA, and are often

detectable in the blood many years before the development of RA [123]. The lead-SNP for PADI2, rs2235910, is genome-wide significant in PBMCs (p-value 6.88×10^{-13}), corroborated by GTEx (v6p [13]; p-value 2.82×10^{-101}); the lead-SNP for PADI4, rs2240335, is an FDR <0.05 *locally*- acting pQTL in PBMCs (p-value 1.40×10^{-5}) but is not a significant eQTL in whole blood. However, it is a significant eQTL – minimum p-value 7.55×10^{-21} – in other (predominantly brain) tissues in GTEx (v6p) [13], thus demonstrating the benefit of measuring cellular protein abundance.

Increased enzymatic activity of PADI2 and PADI4 have previously been found in the synovial fluid of rheumatoid arthritis patients [124,125]. MR demonstrates that increased mean protein level is associated with increased risk of RA for both PADI2 and PADI4 (Table 3.1), leading one to the conclusion that increases in mean PADI2 and PADI4 abundance causally contribute to rheumatoid arthritis. Consistent with this, there have been recent successes in the use of PADI-inhibitors for the treatment of mouse models of rheumatoid arthritis (pan-PADI inhibition [126,127]; PADI4 [128]). Indeed, the pan-PADI inhibitor BB-Cl-amidine has even been shown to reverse immune-mediated joint inflammation [127]. It is currently thought that the association of PADI enzymes and RA is broader than simply generating citrullinated epitopes [128]. PAD4 has previously been shown to affect gene transcription via regulation of the balance between histone arginine methylation and citrullination [129] and, it has previously been shown that PADI2 is also able to translocate to the nucleus and regulate the citrullination of histone H3 [130].

With respect to the other proteins identified as causally contributing to RA (Table 3.1), there is significant pre-existing literature linking HLA subtypes and the extended major histocompatibility complex (MHC) region with RA [131–137]. However, given the strong LD within the region, pleiotropy must be considered carefully when interpreting these results. Finally, RNASET2 has previously been identified as an ethnicity specific ('Asian-specific') RA-associated gene [138].

Table 3.1: Mendelian Randomisation (MR) estimates of the effect of protein on rheumatoid arthritis

SNP:protein estimates from PBMCs, SNP:RA estimate from Okada et al. [53]. A positive β estimate implies that increasing mean protein concentration increases risk of developing rheumatoid arthritis. MAF: Minor allele frequency from the PBMC pQTL results.

Protein	β_{MR} estimate	Standard error$_{MR}$	MR p-value	MAF
HLA-DQA1	-1.33	0.23	1.67×10^{-8}	0.249
NELFE	0.89	0.19	4.93×10^{-6}	0.263
PADI4	1.22	0.31	7.74×10^{-5}	0.317
HLA-B	0.93	0.23	4.47×10^{-5}	0.076
RNASET2	-0.80	0.16	9.85×10^{-7}	0.464
PADI2	0.44	0.10	1.26×10^{-5}	0.351
HLA-F	0.46	0.13	4.04×10^{-4}	0.209

Which PADI? The benefits of breadth.

All known PADI enzymes (PADI 1,2,3,4, and 6) are located in a cluster on human chromosome 1 (Figure 3.2). An assumption of MR is that the SNP (instrumental variable) does not affect the outcome, except via the protein (exposure) itself; or put another way, that the SNP has no horizontal pleiotropy. One benefit of a study of great breadth is that one can directly test this assumption. For example, I have directly measured 3 of the 5 PADI enzymes and, notwithstanding arguments about statistical power, am able to conclude the following: the lead-SNP for PADI2 (rs2235910) is not significantly associated with PADI4 abundance (p-value 0.41), nor was the lead-SNP for PADI4 (rs2240335) significantly associated with PADI2 (p-value 0.66); neither SNP was associated with PADI3 abundance (p-values 0.14 and 0.81, respectively).

Transcription of a genomic region is correlated with the degree to which the DNA CpG sites are methylated. I hypothesised that if a SNP was affecting protein abundance, i.e. it is a pQTL, that it may also be a meQTL for some (or all) of the CpGs across the gene encoding that protein. However, the expected direction of causation is not necessarily clear. I undertook to assess this for PADI2 and PADI4 and present initial encouraging results (Figure 3.2). Obviously, this is a result that will require confirmation and generalisation, however, in these two cases, it is striking. In the case of PADI2, there is an almost exclusive relationship between meQTL significance and physical location within the gene, and in the case of PADI4, almost so. The peak of the PADI4 lead-SNP is slightly broader, and not as tall, this is concordant with the lower beta estimate and less significant p-value of the SNP:protein

association (rs2235910:PADI2 in PBMCs, β estimate 0.19, p-value 6.88×10^{-13}; rs2240335:PADI4 in PBMCs, β estimate 0.11, p-value 1.40×10^{-5}).

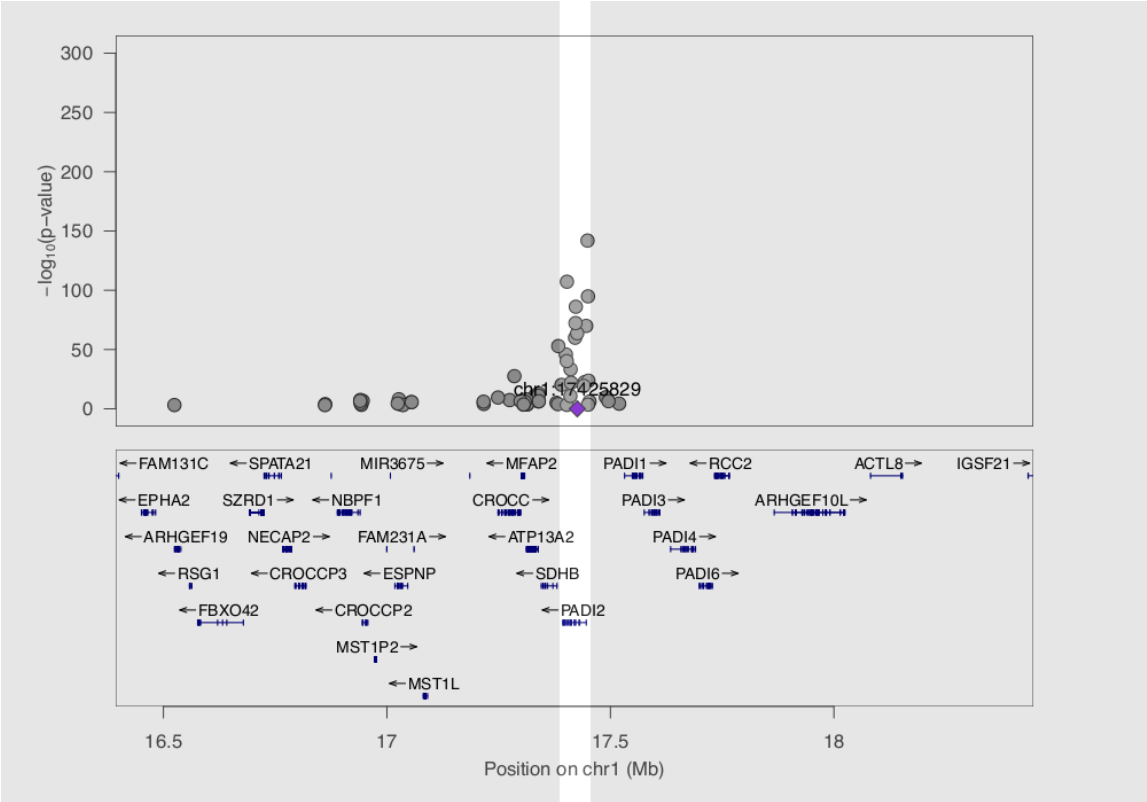
This may represent an efficient way of assessing the pleiotropic effects of a SNP in MR studies of the future. For example, if one were to assess the effect on methylation of the lead-SNP of a known pQTL across the genome, then one might expect a peak of meQTL hits over the gene encoding the protein for which it is a pQTL. However, there may also be other peaks which may represent the pleiotropic effects of the SNP (horizontal or vertical). Also, with the addition of further genetic instruments it should be possible to distinguish these possibilities.

Figure 3.2: Effects of PADI2 and PADI4 pQTL SNPs on CpG methylation

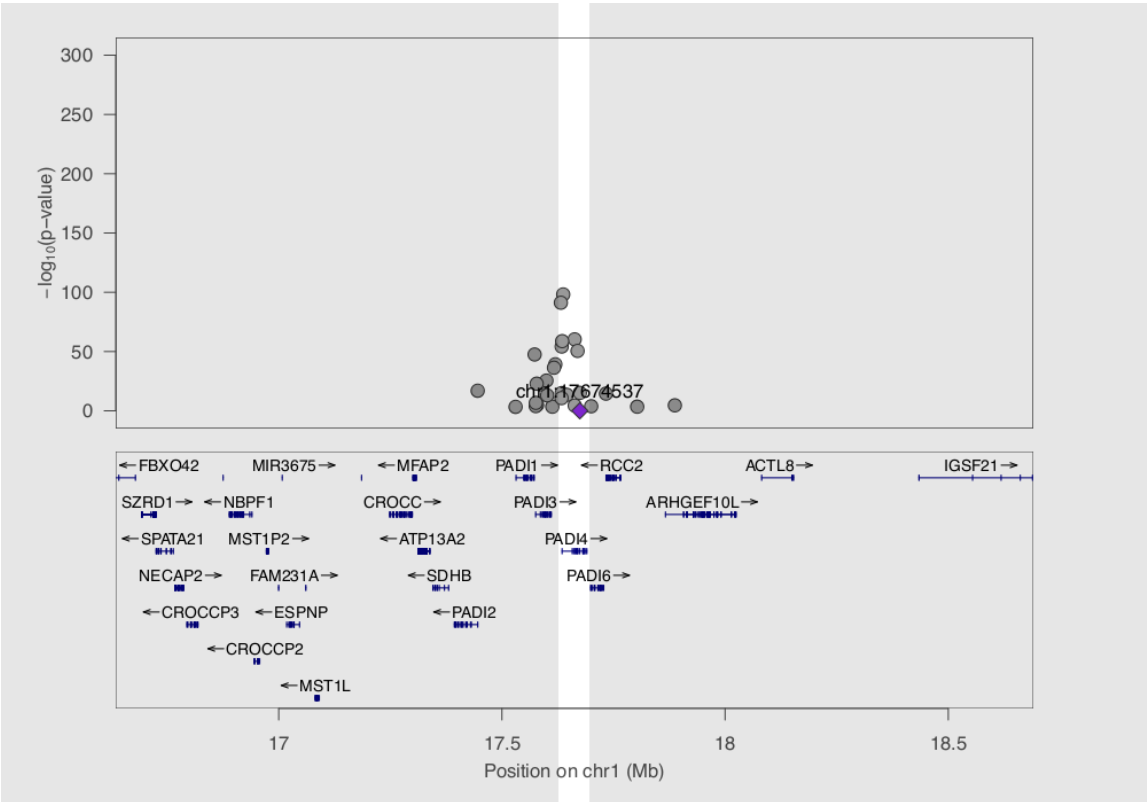
Locus zoom [110] plots of all meQTL (Methods) with an association p-value <$1x10^{-3}$ between the SNP and the CpG. Points on this plot are individual CpG sites. The significance reported is the association of the model SNP:CpG association (as described in Chapter 4). Note, the set of CpG potentially included in each graph is the same. Each plot is centred on the gene encoding the protein for which the SNP is a pQTL with a 1Mb flank. The purple diamond represents the position of the lead-SNP of the pQTL: its position on the y-axis is not meaningful. Shading is a visual aid to mark the gene of interests in the centre of the plot.

A) meQTL associations of lead-SNP for PADI2 (rs2235910). As can be seen the lead PADI2 SNP is associated with CpG methylation changes across the PADI2 gene, with no such enrichment over the PADI1, 3, 4, or 6.
B) meQTL associations of lead-SNP for PADI4 (rs2240335). The lead PADI4 SNP has a broader, less significantly associated, peak: concordant with its lower beta estimate and less significant p-value.

A) PADI2



B) PADI4



Comparison with plasma pQTL

In one of the largest pQTL studies to date, Sun et al. [14] evaluated the plasma proteome to previously unprecedented depth. They assessed 3,608 unique UniProt IDs, of which 1,245 overlap with those assessed in this study. That is, 2,363 UniProtIDs are unique to Sun et al., and 3,734 unique to this study. At first glance, one may be surprised by the degree of overlap observed. However, it is worth noting that protein inclusion in this study is data dependent – that is, the MS features chosen to fragment (and thus attempt identification) are defined by the peptide content of the sample itself – whereas the proteins included by Sun et al. are those chosen *a priori* for inclusion on the panel by SomaLogic.

Neither PADI2, nor PADI4 were assayed in the plasma proteome of Sun et al. [14]. Of the 108 significant (FDR <0.05) *locally*- acting pQTL identified in this study, 37 of the Uniprot IDs were assessed in Sun et al., and of those, 16 were reported as having a *locally*- acting pQTL (within ±1Mb of the transcription start site of the corresponding protein-coding gene). Interestingly, for those UniprotIDs where both studies reported a significant (FDR <0.05 here; p-value $<1.5\times10^{-11}$ Sun et al.) *locally*- acting pQTL, the linkage disequilibrium (r^2) between the lead-SNPs was >0.8 in only 50% (8/16) of cases (13 r^2 values reported: mean 0.50; range 1.21×10^{-3} to 1.0; values 0.00, 0.00, 0.01, 0.01, 0.09, 0.12, 0.68, 0.81, 0.92, 0.93, 0.95, 0.97, 1.00. Lead-SNP identical in 2 cases. 1 pair not assessed as SNP flagged as multiallelic in the LD reference). This disparity may be due to a number of reasons: 1) the SNP density of our genotyped SNPs when compared to the imputed data of Sun et al. may necessitate the use of poor genetic proxies, 2) there may be technical issues with the protein specificity of either of the assays, or 3) there may be truly different molecular

mechanisms underlying the protein abundance change observed in each study. For example, one SNP may increase the rate of intra-cellular breakdown of a protein, without affecting the amount excreted by the cell, or the effects of a SNP may by context dependent, and only affect protein concentration in PBMCs with the effect on plasma abundance as a whole being negligible. Therefore, there is clear virtue in the assessment of cellular proteomes and one should not simply rely on studies of plasma proteins alone to evaluate the genetic control of the proteome in general terms.

Discussion

Here I have presented the results of 251 peripheral blood mononuclear cell samples; representing the first phase of 861 currently being processed. I have examined the genetic associations of 4,433 (FDR <0.01) proteins, each measured in the peripheral blood mononuclear cells of 100, or more, samples. Already, this is the largest population-level study of the cellular proteome to date. Indeed, it is also the deepest assessment of the peripheral blood mononuclear cell proteome to date [111].

Protein, not RNA, is key for drug discovery

eQTL are of great interest and their ease of measurement has made them an attractive early route through which to annotate the pathway between genotype and disease. Based on data from the GTEx project, nearly 50% of common genetic variants were found to be associated with gene expression [13,139]. Unfortunately, without careful interpretation, this apparent success can easily lead to false conclusions regarding the consequences of a given

genetic variant. A SNP, or variants in linkage disequilibrium with it, may affect more than one gene (horizontal pleiotropy), and regulation can occur at all levels of protein production. In addition to this, previous studies have shown pQTLs are enriched for disease linked variation when compared to eQTLs [19]. When comparing the correlation of beta estimates of our PBMC pQTLs with the eQTLs of whole blood from GTEx (v6p) [13], with a co-efficient of determination of 0.45, our results are broadly consistent, with those of previous studies [19]. The biological component contributing to the imperfect correlation between RNA and protein abundance may have important consequences, especially for drug discovery, as the molecular targets of most drugs in use today target protein, not RNA.

Comparison to plasma proteomes and other technologies

The distinction between the cellular and the extra-cellular proteome profile is an important one. The true source of many of the proteins contained within the plasma is far from clear. The plasma proteome represents a composite measure of many cell-types and tissues. I find an incomplete overlap between the lead-SNPs for approximately half of the proteins for which a significant (FDR <0.05 PBMC pQTL; p-value $<1.5x10^{-11}$ plasma pQTL) exists in both PBMCs and plasma. The potential reasons for this are multitudinous. Ranging from proteolytic cleavage of cell-surface receptors [94], through to epitope changes for both antibody and aptamer-based technologies.

One potential pitfall of using mass-spectrometry based assessment of protein abundance is the difficulty in mapping ‘features’ to peptides. A mass-spectrometry proteomics ‘feature’ is

a species that elutes off the column at a specific time with a specific m/z profile. The protocol used to prepare the samples for mass-spectrometry is designed so that each of these features represents a peptide from the digest of the sample. However, whilst all the features (when present) are measured in all the samples, without MS2 fragmentation of the feature, it is not clear which peptide a particular feature represents. However, the identification of features in one sample carries to other samples (i.e. 'match-between-runs'), and so the more samples run (i.e. the more MS2 spectra acquired) and the use of a high-pH, reversed-phase, peptide library greatly increases the number of features identified per sample. A high-pH reversed-phase peptide library is designed to separate peptides by hydrophobicity in an orthogonal manner to the low-pH reversed-phase gradients used as part of the LC-MS/MS setup. Not only this, some of the additional peptides identified are likely to derive from previously unidentified proteins, and so, as more features are identified, the depth to which the proteome is covered will also increase.

As is standard in the mass-spectrometry proteomics field, in this study I used a database method involving *in silico* protein digestion. Non-synonymous, or splicing changes are not represented in the database, leading to spuriously low values for proteins for which this is an issue. Future work focussing on the features themselves, peptides directly, or a mutation / splice variant aware search are all potential solutions to this problem. However, in spirit, this issue is not dissimilar to that of altered epitopes (due to any non-synonymous variant that changes the conformation of the protein in such a way as to change the epitope) that plagues antibody and aptamer-based assays, as mentioned above. Notwithstanding the above discussion regarding the advantages of pQTLs over eQTLs for drug discovery, the

observed overlap of our results with eQTLs from GTEx (v6p) [13] – identified through an orthologous method – does provide a degree of confidence in the pQTLs presented here. In mass-spectrometry based proteomics, as peptide (and hence protein) identification is based on the HPLC elution and m/z profile of a feature, mass-spectrometry proteomics is not limited to a set of proteins that has been arbitrarily chosen to appear on an assay, but by the resolution of the HPLC and mass spectrometer themselves. The breadth of coverage of the proteome, and the post-translational modifications of such, by mass-spectrometry proteomic methods is truly spectacular. It enables a more thorough assessment of the effects of a variant on all the genes encoding proteins within a genomic locus than has ever been possible before.

Rheumatoid arthritis

I report significant MR links between PADI2 and PADI4 with RA, as well as other previously documented results. PADI inhibitors have shown great promise in the prophylaxis and treatment of RA models in mice [126–128]. In addition, I demonstrate a link between a pQTL and the meQTL overlaying the gene of the protein under study (Figure 3.2). How ubiquitous this relationship is, and whether it extends to *distantly*- acting pQTL remains to be seen, and is an area of potential future work. However, it does present the exciting possibility of assessing the likely pleiotropic effects of a given SNP at a locus.

Future work

First and foremost, data collection of the remaining 610 samples needs to be completed, this is due to be completed by November 2019. Following the acquisition of the data, it will require to be searched. At which point, the analyses above will need to be redone. The difficulty of the peptide search and these analyses should not be underestimated. Potential difficulties include the computational resources required to perform a peptide search on 681 samples plus a peptide library. At present we have fit the mixed model, with the SNP, in one go. This may not be practical when the sample size is increased 3.4-fold, the number of proteins identified increases by approximately 50%, and imputed genotype data is used. Once data acquisition is complete, full assessment of the correlation structure of the proteins identified per individual and the identification of outliers should be considered. Other potential improvements to the analysis include the investigation of other potential protein / peptide imputation strategies, for example, k-nearest neighbours' imputation, and the analysis of unique peptides only. In the present analysis, we co-vary only for haemoglobin concentration, as a proxy for the degree of erythrocyte contamination in the sample. However, in the fullness of time it is likely to become possible to co-vary for a number of other cell-types by including more cell-surface receptors as proxies. In this way it would be possible to compensate for heterogeneity introduced between samples because of cell-type composition differences. When the analysis is run using a larger sample size, and the GWA run using imputed data, standard heterogeneity testing, for example HEIDI testing [27], as discussed in Chapter 2, is likely to become useful. However, the investigation of the use of meQTL as a marker of horizontal pleiotropy is an exciting avenue to pursue in this arena also.

When using HPLC-MS/MS data, it is worth mentally separating the analysis into two compartments, that of the associations of HPLC-MS/MS features, for example SNPs and disease, and that of the identification of the features themselves. There is no theoretical reason, other than burden of multiple testing, not to perform a GWA on each of the features detected, with the perspective of identifying the molecule underlying the feature in a subsequent – targeted – HPLC-MS/MS study. This analysis method would also have the secondary benefit of blinding the researcher to the exposure variable, thereby removing a potential source of bias.

Conclusions

In this exploratory analysis, I have clearly demonstrated the feasibility of assessing the cellular proteome of peripheral blood mononuclear cells. I present the result of the first 251 samples, of a total of 861 due to be processed. The sample preparation of the 861 has been completed and data acquisition of the remaining 610 is currently underway. However, I have already obtained the most in-depth profile of the proteome of peripheral blood mono-nuclear cells ever completed. As sample size increases, and with the inclusion of a high pH reversed-phase peptide library, the number of peptides and proteins identified will only increase. The large increase in sample size will also enable far more nuanced analyses, including, but not limited to, the assessment of post-translational modifications.

4. DNA CpG methylation

Introduction

Cytosine methylation of DNA is an important epigenetic modification in Eukaryotes and, when disturbed, has been associated with the pathogenesis of numerous diseases [12,20,21]. In molecular terms, it is the covalent addition of a methyl group to the DNA base cytosine, resulting in 5-methylcytosine. In humans, this mark predominantly occurs at CpG dinucleotides and, outside of CpG islands (where the situation is more complex), is wide-spread across the genome [23].

The degree to which a given CpG site is methylated (across a population of cells, and between individuals) can be under genetic control. Genetic sequence variation, and the surrounding variants in linkage disequilibrium with it, that lead to alteration of the methylation status of a particular CpG are termed a methylation quantitative trait locus, or meQTL. These can be local to the genomic location of the CpG site itself, or distant to it. Throughout this chapter, when describing the results I generated (unless explicitly stated otherwise) I will use *'inter-chromosomal*' exclusively to describe the situation where the SNP and CpG are located on different chromosomes, *'locally*- acting' when the SNP:CpG pair are located within 1Mb of each other (on the same chromosome), and 'unconstrained' where neither such restriction had been applied.

The specific molecular mechanisms that enable a genetic sequence variant to influence DNA CpG methylation are not well understood. At certain genomic loci, CpG methylation is a

dynamic process and is associated with a number of key factors. These include variable transcription factor binding and transcription of a region, as well as changes to the local chromatin state [23]. It has previously been shown that the affinity of transcription factors for DNA can be both sequence and methylation dependent [140,141]. The links between transcription factor binding-site occupancy, transcription, and epigenetic changes (including both CpG methylation and chromatin modification) are extremely complex [142,143] and are currently only partially understood.

As with the prevalence of CpG methylation, transcription factor occupancy at transcription factor binding sites can also be under local and distant genetic control. A systematic difference in the mode of inheritance between genetic variation *cis*- and *trans*- (distinguished based on allele specificity) to transcription factor binding sites on transcription factor occupancy has recently been shown (in mouse livers) for three transcription factors: FOXA1, CEBPA, and HNF4A [143]. They found that genetic variation c*is*- to transcription factor binding sites was more likely to demonstrate an additive mode of inheritance, whereas genetic variation *trans*- to transcription factor binding sites was more likely to exert dominance control.

In addition to the differences in the mode of inheritance of *cis*- and *trans*- to transcription factor binding sites, there is a known enrichment of meQTL target CpG sites in upstream and 5' untranslated regions [12], suggesting a relationship between variable methylation and regulatory control of transcription. Given this evidence, and the intimate link between methylation and transcription factor binding site occupancy [144], I hypothesised that variable

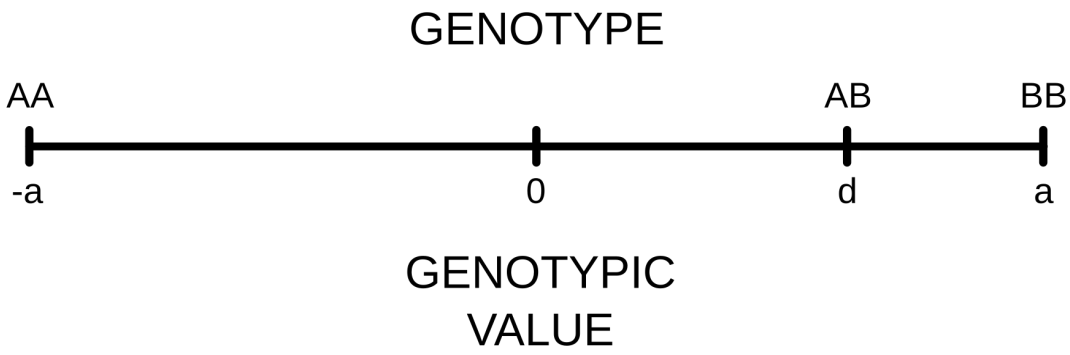
DNA CpG methylation in the binding sites of many human transcription factors may be under dominance genetic control. Transcription factors are the canonical *trans*- acting factors, and I therefore focused the majority of this chapter on a confident subset of *trans*-acting meQTL: inter-chromosomal meQTL.

Mendelian dominance describes the situation where a recessive allele is masked by a dominant one, that is, a heterozygote will express the phenotype of the dominant allele over the recessive one. To put this another way, dominance can be thought of as the deviation of the heterozygote phenotype from the midpoint of the phenotype of the two homozygotes. Dominance is frequently seen in single gene disorders, for example, myotonic dystrophy, Huntington's disease, and Marfan's syndrome [145]. However, for 'complex' traits, classical GWA studies have focussed on estimated allelic effects rather than attempting to delineate additive and dominance genetic effects on trait value. Using this model, these studies have been remarkably successful in finding disease associated variants. However, both additive and dominance components contribute to an allele's effect, and one may expect dominance deviation when variation acts in *trans*-. This is because effects due to a diffusible mediator are unlikely to be chromosome specific at the distant site. Approximately one-third of trait-associated SNPs examined in one study were identified as affecting distant DNA CpG methylation by Bonder et al. [15]. Therefore, as I focussed on inter-chromosomal genetic control of DNA CpG methylation, I fit a model explicitly including both additive and dominance genetic effects.

One method to assess *trans*- acting effects and their targets (e.g. transcription factor binding sites) is to use a molecular phenotype, such as DNA CpG methylation, that provides genome-wide coverage of targets. Unfortunately, previous studies attempting to analyse *trans*- acting genetic effects at scale and/or dominance effects have generally been underpowered. In the case of *trans*- acting control, power is reduced, in part, simply due to the greatly increased burden of multiple testing when including the remainder of the genome, rather than just the local region. In general terms, it is more difficult to identify dominance genetic effects compared to additive effects because of the construction of the model: additive effect estimates are based on the difference between the two homozygotes, i.e. 2 'a', whereas the dominance effect is the difference of the heterozygote from the midpoint of the two homozygotes, i.e. 'd' (Figure 4.1, Equation 4.1). In addition, if attempting the genome-wide coverage of both the genetic variants and the molecular phenotype, the number of tests increases in line with the product of independent variants and phenotypes. We present a well-powered, GWA study of DNA CpG methylation at 573,027 CpG sites of 5,101 individuals from a single cohort measured using the Illumina EPIC array and assess additive and dominance genetic contributions simultaneously. Using these data, we analyse the molecular intermediates mediating the effect of inter-chromosomal SNPs and review enrichments at their targets.

Figure 4.1: Additive / Dominance genotypic values.

Adapted from [146]. The zero point on the line is the midpoint between the two homozygotes. Genotypic value is arbitrarily assigned as deviations from this value: the two homozygotes as ‘a’ and ‘-a’, and the heterozygote as ‘d’. Note that the sign of ‘a’ is dependent upon which allele (‘A’ or ‘B’) is coded as the effect allele in the regression, the sign of ‘d’ is not. In this case, if ‘BB’ represents the more methylated homozygote, the genotypic value of ‘AB’ is closer to that of ‘BB’ and is therefore more methylated than the midpoint of the two homozygotes: allele ‘B’ is displays dominance over allele ‘A’.



Methods

Generation Scotland

Generation Scotland: Scottish Family Health Study (GS) is a population- and family-based cohort from the Scottish population collected between 2006 and 2011 and has been described in detail elsewhere [114].

All participants provided written informed consent and ethical approval was provided by the East of Scotland Research Ethics Service committee on research ethics (REC references 15/ES/0040).

Genotyping

Samples for DNA extraction (blood, or occasionally saliva) were obtained at the time of recruitment. Genotyping was carried out by the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland using the HumanOmniExpressExome-8 v1.0 or v1.2 BeadChips with Infinium chemistry (Illumina). Genotypes were processed using the GenomeStudio Analysis software v2011.1 (Illumina) and called using BeadStudio-Gencall v3.0 (Illumina). The details of blood collection and DNA extraction are also provided elsewhere [115].

Subsequent quality control removed individuals with <98% call rate, SNPs with <98% call rate, and SNPs with Hardy-Weinberg equilibrium p-value $<1x10^{-6}$. After initial quality

control, 604,858 genotyped autosomal SNPs remained. The genotyped data were imputed utilising the Sanger Imputation Service to the HRC panel v1.1, as described previously by Nagy et al. [147]. The data were pre-phased using SHAPEIT v2.r873 [148] + duohmm12 [103] and imputed with PBWT [149].

Measurement of methylation

Whole blood genomic DNA (500ng) was treated with sodium bisulfite using the EZ-96 DNA Methylation Kit (Zymo Research) and DNA methylation was assessed using Illumina Infinium MethylationEPIC BeadChip technology (Illumina), as per the manufacturer's instructions. The arrays were scanned using an Illumina HiScan scanner (Illumina) and initial inspection of array quality was carried out using Illumina GenomeStudio Analysis software v2011.1 (Illumina).

Quality control of the DNA methylation data was carried out before normalisation. The R package shinyMethyl [150] was used for preliminary quality control. This quality control step removed 81 samples, based on the following criteria 1) overall array signal intensity and control probe performance outliers, 2) samples with a mismatch between recorded gender and predicted gender based on X and Y chromosome DNA methylation, and 3) genetic ethnic outliers for the cohort identified by principal component analysis [151]. Further quality control was performed using the 'pfilter' function in the R package watermelon [152], samples were removed if ≥1% sites had a detection p-value of >0.05. This removed 18 further samples. Finally, this left 5,101 samples for further processing. Before normalization,

individual probe-sample pairs with a detection p-value of >0.05 were removed. Normalization was performed using function 'preprocessNoob' in the R package minfi [153].

In order to remove potential technical confounders, linear mixed modelling was used to pre-correct each probe. This model included the following fixed effects: top 50 principal components of control probe intensities (which explained 99% of variation in control probe intensities), appointment clinic centre, processing batch, year of the visit, and Sentrix position (position of the sample in Illumina slide); and the random effects: appointment date and Sentrix ID (Illumina slide). The model converged successfully for 712,595 sites, and the resultant residualised M-values were used as DNA methylation phenotypes in downstream analysis. For individual sites, outlier samples with residualised-M-values more than five interquartile ranges from the nearest quartile were removed.

Finally, we fit biological covariates to create residuals for GWA. Cell-type proportions were estimated for granulocytes, monocytes, B-lymphocytes, natural killer cells, CD4+ T-lymphocytes and CD8+ T- lymphocytes using the 'estimateCellCounts' function in R package minfi [153]. The following mixed linear model was fit: Fixed effects) age, age^2, gender, cell-type proportions for granulocytes, B-lymphocytes, natural killer cells, CD4+ T-lymphocytes and CD8+ T-lymphocytes, season of the visit, appointment time of the day, appointment day of the week; Random effects) genomic relationship matrices, G (genomic relationship matrix) and K (kinship relationship matrix), and three environmental relationship matrices, F (environmental matrix representing nuclear-family-member relationships), S (environmental matrix representing full-sibling relationships) and C (environmental matrix representing

couple relationships). The resulting residuals were inverse rank transformed prior to GWA analysis in a simple linear model. Data preparation and residualisation have been described previously [2] but are included here again for completeness.

Genome-wide association

Model

A combined, additive and dominance, model was fit to the dataset (Equation 4.1; Figure 4.1). Individuals were randomly split into a discovery set of 4,101 (set 1), and a replication set of 1,000 individuals (set 2). This model was fit to all CpG sites for which the residualisation had successfully completed (n = 573,027).

Equation 4.1:

$$y_i = \beta_{intercept} + \beta_{ADD} x_{ADD_i} + \beta_{DOM} x_{DOM_i} + \varepsilon_i$$

- y is the residualised methylation values (above).
- $\beta_{intercept}$ is the intercept term.
- β_{ADD} and β_{DOM} are the additive and dominance effect-size estimates.
- x_{ADD_i} is, for individual *i*, the sum of the minus probability of the genotypes AA plus the probability of the genotypes BB: -p(AA)+p(BB).
- x_{DOM_i} is, for individual *i*, the probability of genotypes AB: p(AB).
- ε_i is the error.

The model was fit to those SNPs that had total cumulative genotype count of >1 per genotype (N.B. subsequently filtered to ensure a total cumulative genotype count of >10 per genotype). Results for which the model p-value was $<1x10^{-3}$ were retained.

Post-GWA processing

Maximal independent set

The pairwise Pearson's correlation coefficient of all CpG passing all quality control measures (n = 573,027) was calculated using the processed residualised phenotype data. A maximal independent set was calculated so that no pair of CpG in the resulting graph had a squared-correlation coefficient >0.2. 512,601 CpG were present in the reduced set.

LD clumping

Genome-wide association results for CpG sites passing all quality control measures (n = 573,027), for which a genomic location was reported in the Illumina EPIC array manifest (B3) (Illumina) present in the maximal independent set of CpGs are reported. Genetic variants were filtered to include only those with an imputation info score (SNPTEST v2) of >0.95. Per CpG, SNPs were clumped by linkage disequilibrium. That is, a lead variant (smallest p-value in the genome remaining un-clumped was selected) and variants on the same chromosome were clumped together if the squared Pearson's correlation coefficient between the two was greater than 0.2. This process was repeated iteratively until all SNPs had been assigned to a clump. An unrelated set of European populations (CEU, FIN, GBR, IBS, and TSI) from the 1,000 Genomes project [92,93] were used as the reference population for this procedure.

Finally clumps for which the lead variant had fewer than 10 of each genotype were removed.

CpG and LD clumping (pre-regression analysis)

Just as one can extend the concept of a 1D vector to a 2D matrix, the same is true for a standard Manhattan plot and the results presented here. That is, literally and figuratively, genome-wide methylation data adds an additional dimension to standard SNP data. If a standard Manhattan plot is a 1D vector of $-\log_{10}$(p-values), the stack of CpG Manhattan plots, one per CpG, can be thought of as a 2D matrix of $-\log_{10}$(p-values). Unfortunately, the SNPs are not independent of each other, and neither are the CpG. Genetic variants were filtered to exclude the required window: unconstrained analysis – SNP to CpG distance ≥50bp (in order to exclude technical variation); inter-chromosomal – SNP and CpG located on different chromosomes. SNPs with an imputation info score (SNPTEST v2) of ≤0.95, and those pairs for which the SNP was in LD (r^2 >0.2) with any SNP within 10bp of the CpG site were also excluded. For each CpG site, the SNP with the largest F-statistic was then extracted. A process akin to LD clumping of SNPs was applied to the CpGs (Algorithm 1), and then standard LD clumping (Algorithm 1, but using SNPs and SNP correlations instead of CpGs and CpG correlations).

These filtering steps result in a list of (quasi-) independent SNP:CpG pairs. An unrelated set of European populations (CEU, FIN, GBR, IBS, and TSI) from the 1,000 Genomes project [92,93] was again used as the LD reference. Finally clumps for which the lead variant had fewer than

10 of each of each genotype were removed. Note that this is a very conservative analysis for the following reasons: 1) per CpG we take the maximum f-statistic for the model amongst all SNPs, therefore the number of SNPs included per CpG is limited to one, 2) the clumping of CpGs and SNPs were applied in series rather than together, and 3) minimum genotype count filtering was applied post-clumping. However, despite this, we still detect many significant SNP:CpG pairs.

We performed the above separately for the discovery (set 1; n = 4,101) and replication (set 2; n = 1,000) sets.

Algorithm 1:

1. List SNP:CpG pairs.
2. Select the most significant (largest F-statistic) SNP:CpG pair yet to be clumped as the lead pair for a new clump and remove it from the list.
3. Remove from the list, all SNP:CpG pairs for which the CpG is correlated (squared Pearson's correlation coefficient) with the CpG of the lead pair with a value greater than 0.2.
4. Repeat until the list is empty.

Enrichment analysis using String [154] v11

SNPs with a significant (Bonferroni correction: p-value $<5x10^{-8}/512,601$) association (model) with any CpG (in the CpG and LD clumped results) were selected. This was performed for

both the unconstrained and inter-chromosomal sets. Significant *locally*- acting eQTL co-localising with these SNPs were selected from the 'whole blood' results of GTEx (v6p) [13], the mapping window of which was limited to ±1Mb from the transcription start site. These Ensembl Gene IDs were converted to HGNC symbols using Ensembl Biomart (GRCh37) [155] [accessed 25/06/2019] and input into String v11 [154]. 'Gene type' was extracted from, and defined by, Ensembl.

I assessed: 1) the network of inter-chromosomal meQTLs (Bonferroni correction, $5x10^{-8}/512,601$) under significant genetic control from the discovery set, 2) the network of inter-chromosomal meQTLs (Bonferroni correction: p-value $<5x10^{-8}/512,601$) under significant genetic control from the replication set, and 3) two randomly selected sets of 1,000 SNPs from GTEx (v6p) [13] and their associated eQTL. The randomly generated sets were used for comparison to ensure that the enrichment observed was not simply due to enrichment within GTEx itself, rather than that associated with the inter-chromosomal meQTLs. We considered enrichments in 'Biological Process' (GO) [156], 'Molecular function' (GO) [156], 'Cellular Component' (GO) [156], 'Reactome Pathways' [157], 'UniProt Keywords' [158], 'PFAM Protein Database' [159], 'INTERPRO Protein Domains and Features' [160], and 'SMART Protein Domains' [161] as output by String.

Mapping to all transcription factors

SNPs with a significant (Bonferroni correction: p-value $<5x10^{-8}/512,601$) association (model) with any CpG (inter-chromosomal; discovery set; CpG and LD clumped results) were

selected. Significant eQTL co-localising with these variants were selected from the 'whole blood' results of GTEx (v6p) [13], as above, and the Ensembl Gene IDs of those classified as 'protein coding' in Ensembl (GRCh37), matched to all human transcription factors as identified by Lambert et al.[30]. Note that ENSG00000250312 (ZNF718) was misclassified in Ensembl GRCh37 [155] [accessed 25/06/2019] as a lincRNA, this has subsequently been updated in Ensembl GRCh38 [162] [accessed 24/10/2019]. ZNF718 has been included in the results as an addendum.

Effect of increased trans- acting factor RNA on DNA CpG methylation at distant sites

Taking the list of CpGs and LD clumped inter-chromosomal results, for each of the significant (Bonferroni correction, p-value $<5x10^{-8}/512,601$) meQTL that co-localised with a *locally*- acting eQTL in whole blood from GTEx (v6p) [13], following effect allele harmonisation and removal of duplicate SNP and Ensembl gene IDs, linear regression analyses were performed as per Equation 4.2 and Equation 4.3. The construction of Equation 4.2 is to answer the question: is there an association between the allele that increases RNA abundance (*locally*- acting) and increases methylation on a different chromosome (inter-chromosomal); and Equation 4.3 is such as to answer the question: is there an association between the dominance effect estimate on methylation of a different chromosome (inter-chromosomal), orientated to the methylation status of the homozygote of the effect allele, and the estimate of the effect allele on the transcript abundance (*locally*- acting).

Equation 4.2: meQTL additive beta estimate against eQTL beta effect estimate

$$\beta_{ME_{ADD}} \sim \beta_{RNA} + \text{intersect}$$

- $\beta_{ME_{ADD}}$ is the additive beta estimate from the meQTL study, as per Equation 4.1.
- β_{RNA} is the eQTL beta effect estimate (estimated allelic effect) from GTEx (v6p) [13].

Equation 4.3: meQTL dominance beta estimate against eQTL beta effect estimate

$$\beta_{ME_{DOM}} * sign(\beta_{ME_{ADD}}) \sim \beta_{RNA} + \text{intersect}$$

- $\beta_{ME_{DOM}}$ is the dominance beta estimate from the meQTL study, as per Equation 4.1.
- $sign(\beta_{ME_{ADD}})$ is the sign of the additive beta estimate from the meQTL study, as per Equation 4.1, coded as -1 and 1, for negative and positive.
- β_{RNA} is the eQTL beta effect estimate (estimated allelic effect) from GTEx (v6p) [13].
- Note that the asterisk is a multiplication, not an implicit interaction set in the regression.

Regression of additive and dominance effect estimates

Taking the list of CpG and LD clumped results, significant (p-value < $1x10^{-13}$) inter-chromosomal meQTLs, we fit β_{DOM} (dominance effect estimate) and $|\beta_{ADD}|$ (absolute magnitude of additive effect estimate) against a base-model (Equation 4.4 and Equation 4.5) including: the other of β_{DOM} and $|\beta_{ADD}|$ plus the frequencies of the heterozygote and minor homozygote, the SNP and CpG chromosomes (as a categorical variable), the minor allele type (as a categorical variable), and whether the CpG is located within a CpG island (based on the UCSC [163,164] [accessed 23/06/2019] track 'cpgIslandExt'), and the SNP or CpG annotation in question based on its genomic position (Equation 4.4 and Equation 4.5).

Annotations, as provided as part of the LOLA [165] package, were used from Cistrome [166] – epigenome (histone marks), Codex [167] (transcription factors), and Encode segmentation ('wgEncodeAwgSegmentation') [163,164,168]. Annotations were fit one at a time, to both SNP and CpG location, and considered to be significantly associated if they passed a, within collection, Bonferroni correction. We fit the model within the discovery set and attempted to replicate those that reached Bonferroni significance in the discovery set.

Equation 4.4: Magnitude of dominance effect regression

$$\beta_{DOM} \sim |\beta_{ADD}| + \mathrm{Chr}_{SNP} + \mathrm{Chr}_{CpG} + \mathrm{Allele}_{Minor} + \mathrm{Freq}_{minor\ homozygote} + \mathrm{Freq}_{heterozygote} + \mathrm{CpGisland} + \mathrm{annotation} + \mathrm{intersect}$$

Equation 4.5: Absolute magnitude of additive effect regression

$$|\beta_{ADD}| \sim \beta_{DOM} + \mathrm{Chr}_{SNP} + \mathrm{Chr}_{CpG} + \mathrm{Allele}_{Minor} + \mathrm{Freq}_{minor\ homozygote} + \mathrm{Freq}_{heterozygote} + \mathrm{CpGisland} + \mathrm{annotation} + \mathrm{intersect}$$

- β_{DOM} is the effect-size estimate for dominance, as defined in Equation 4.1.
- $|\beta_{ADD}|$ is the absolute magnitude of the effect-size estimate for additivity, as defined in Equation 4.1.
- Chr is chromosome, the subscript defines which location it refers to, SNP or CpG, dummy coded.
- $\mathrm{Allele}_{\mathrm{Minor}}$ is the minor allele type, dummy coded.
- Freq is the genotype frequency of the minor homozygote and heterozygote, as referred to in the subscript.

- CpGisland is the CpG within a CpG island as defined in UCSC [163,164] [accessed 23/06/2019] track 'cpgIslandExt'), binary variable.

Results

Methylation quantitative trait loci are abundant throughout the genome

We demonstrate widespread genetic control of CpG methylation (Figure 4.2), including both additive, and dominance genetic control of CpG methylation (Table 4.1). In a maximal independent set of CpGs (squared-correlation <0.2 between any pair of CpG sites), following LD clumping (Methods) we identify 683,842 significant (Bonferroni correction for multiple testing) SNP:CpG pairs: representing significant genetic control of 177,032 CpGs (34.5%; n = 512,601), and 442,218 SNPs in the discovery set (n = 4,101). Of these, 262,816, 110,527, and 204,141, respectively, replicate (n = 1,000; Table 4.1).

Table 4.1: The number of SNP:CpG pairs in each meQTL set.

Bonferroni correction for multiple testing used throughout, see main text for full description. When listed as discovery / replication, the SNP:CpG pairs that were found to be significant in the discovery set were looked up in the replication set, and the significance threshold determined by the number of significant results in the discovery set. However, when listed as 'set 1' and 'set 2', the two sets were assessed independently, with Bonferroni thresholds determined accordingly.

Set	**Size (discovery / set 1)**	**Size (replication / set 2)**
LD clumped, minimum independent CpG set, model significant (discovery / replication)	683,842	262,816
LD clumped, minimum independent CpG set, model significant (discovery / replication), additive control (set 1 / set 2)	555,928	190,005
LD clumped, minimum independent CpG set, model significant (discovery / replication), dominance control (set 1 / set 2)	11,143	3,137
LD and CpG clumped, *locally*- acting genetic control (set 1 / set 2)	43,143	26,001
LD and CpG clumped, *locally*- acting additive genetic control (set 1 / set 2)	40,421	25,117
LD and CpG clumped, *locally*- acting dominance genetic control (set 1 / set 2)	5,562	1,990

LD and CpG clumped, inter-chromosomal genetic control (set 1 / set 2)	3,364	1,333
LD and CpG clumped, inter-chromosomal additive control (set 1 / set 2)	3,161	1,294
LD and CpG clumped, inter-chromosomal dominance control (set 1 / set 2)	302	122

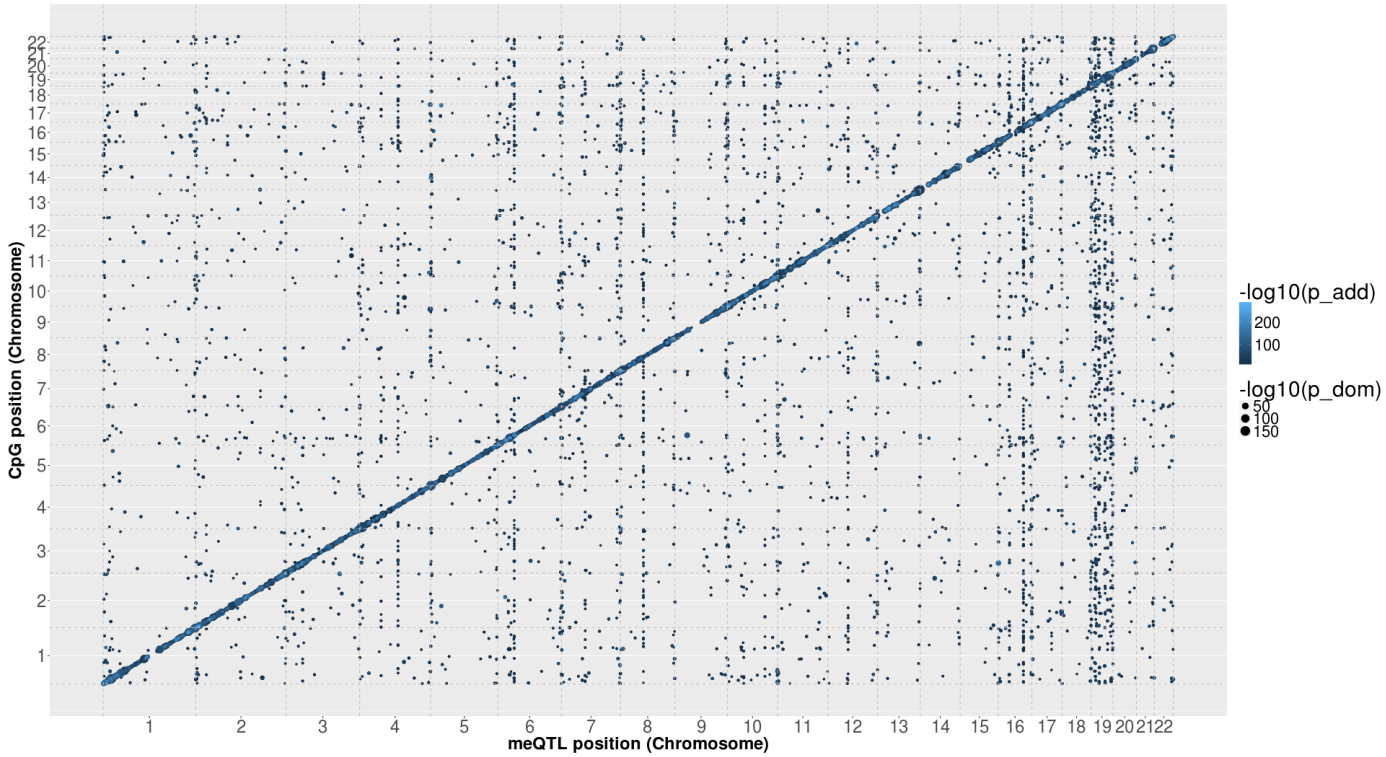
Of the 683,842 (discovery), and 262,816 (replication) SNP:CpG pairs under significant (Bonferroni correction: p-value <$5x10^{-8}$/512,601, and <0.05/683,842, respectively) genetic control, 555,928, and 190,005 demonstrate significant additive control (Bonferroni correction: p-value <0.05/683,842, and <0.05/262,816), and 11,143 and 3,137 demonstrate significant dominance control (Bonferroni correction, <0.05/683,842, and <0.05/262,816), respectively. Note that the replication set is approximately 4-times smaller than the discovery set.

Locally- acting genetic control of DNA CpG methylation is very common

Locally- acting genetic control of CpG methylation is very common, as can be seen from Figure 4.2. The diagonal line on this figure represents independent (LD clumped; Methods) meQTLs for which the genomic location (chromosome, position) of the SNP is close to the CpG location. Following CpG and LD clumping, we identified 43,143 significant *locally*-acting (SNP to CpG distance between 50bp and 1Mb) meQTLs in set 1, and 26,001 in the set 2 (Bonferroni correction: p-value <$5x10^{-8}$/512,601). Of these, 93.7% and 96.6% exert significant (40,421 and 25,117; Bonferroni correction: p-value <0.05/43,143 set 1, and <0.05/26,001 set 2) additive genetic control, respectively; and 12.9% and 7.7% exert significant (5,562 and 1,990; Bonferroni correction: p-value <0.05/43,143 set 1, and <0.05/26,001 set 2) dominance control (Table 4.1).

Figure 4.2: Aerial view of Manhattan (plot)

Plot of SNP genomic location to CpG genomic location for all significant, replicated (Bonferroni correction to both) meQTLs following LD clumping (Methods).



Inter-chromosomal genetic control of DNA CpG methylation is also widespread

Inter-chromosomal genetic control, by definition, involves at least two distinct loci. Following CpG and LD clumping, I identify, 3,364 significant (0.7%; Bonferroni correction: p-value <$5x10^{-8}$/512,601) inter-chromosomal SNP:CpG pairs (set 1) and 1,333 significant (0.3%; Bonferroni correction: p-value <$5x10^{-8}$/512,601) pairs (set 2). I investigate the potential molecular intermediates below.

Molecular intermediates of inter-chromosomal meQTLs

C2H2 zinc-finger proteins (± KRAB domain)

We delineate the molecular intermediates of inter-chromosomal genetic control of DNA CpG methylation by mapping the meQTLs under significant (Bonferroni correction, p-value <$5x10^{-8}$/512,601) inter-chromosomal genetic control (separately for both set 1 and set 2) to significant *locally*- acting eQTL in whole blood using GTEx (v6p) [13] and performing gene-set enrichment analyses using String v11 [154] (Methods). In general terms, we find strong enrichments for C2H2-ZF proteins, especially those containing a KRAB domain, as well as enrichments for transcriptional processes (Table 4.2). Of those significant in the discovery set, all successfully replicate (Bonferroni correction, p-value <0.05/17), and the odds ratios in the discovery set range from 2.15 to 3.76. As the replication set was analysed in parallel to the discovery set from the point of residualisation onwards, including the application of multiple-testing corrections, it is interesting to note that the odds ratios are uniformly larger, and the p-values smaller, in the replication set. This implies that the *locally*- acting eQTL of the proteins in these sets, are those associated with the most significant inter-

chromosomal meQTL (i.e. those that remain discoverable in the smaller replication set). Clearly, there is correlation amongst the annotations in Table 4.3, however, these broadly encompass the nested sets of KRAB domain containing transcription factors, C2H2 transcription factors, and transcription itself.

Table 4.2: Enrichments amongst the genes of the locally- acting eQTL co-localising with inter-chromosomal meQTL.

Significant enrichments amongst whole blood *locally*- acting eQTL from GTEx (v6p) [13] that co-localise with an inter-chromosomal meQTL, when compared to a randomly chosen set of whole blood *locally*- acting eQTL. Bonferroni correction in discovery based on 9,099 tested annotations. Odds ratios are: odds in eQTL set co-localising with an inter-chromosomal meQTL over the odds in random eQTL set. p-values are Fisher's Exact test p-values. Enrichment information from String v11 [154]. All enrichments from the discovery set replicate successfully. Number of proteins in sets: discovery 590, random background one 798; replication 278, random background two 791.

Category	Term	Description	Discovery		Replication	
			Odds Ratio	p-value	Odds Ratio	p-value
Pfam	PF00096	Zinc finger, C2H2 type	3.43	4.73×10^{-9}	5.61	5.02×10^{-12}
InterPro	IPR036236	Zinc finger C2H2 superfamily	3.13	3.00×10^{-8}	5.76	1.83×10^{-12}
InterPro	IPR013087	Zinc finger C2H2-type	3.04	5.99×10^{-8}	5.76	1.83×10^{-12}
SMART	SM00355	zinc finger	2.96	9.48×10^{-8}	5.35	7.70×10^{-12}
InterPro	IPR036051	KRAB domain superfamily	3.76	2.43×10^{-7}	6.04	2.95×10^{-10}
Pfam	PF13912	C2H2-type zinc finger	3.37	2.63×10^{-7}	5.53	3.82×10^{-10}
InterPro	IPR001909	Krueppel-associated box	3.65	3.54×10^{-7}	5.92	2.33×10^{-10}

Keyword	KW-0804	Transcription	2.20	4.58×10^{-7}	3.11	2.85×10^{-9}
Process	GO:0006351	transcription, DNA-templated	2.15	5.25×10^{-7}	3.11	9.05×10^{-10}
Process	GO:0097659	nucleic acid-templated transcription	2.15	5.25×10^{-7}	3.11	9.05×10^{-10}
Process	GO:0032774	RNA biosynthetic process	2.15	5.25×10^{-7}	3.11	9.05×10^{-10}
Function	GO:0003700	DNA-binding transcription factor activity	2.38	5.45×10^{-7}	3.91	9.09×10^{-11}
Function	GO:0000981	DNA-binding transcription factor activity, RNA polymerase II-specific	2.37	6.42×10^{-7}	3.99	1.56×10^{-10}
SMART	SM00349	krueppel associated box	3.24	7.32×10^{-7}	5.45	2.93×10^{-10}
Function	GO:0140110	transcription regulator activity	2.25	7.59×10^{-7}	3.25	3.39×10^{-9}
Pfam	PF01352	KRAB box	3.25	1.42×10^{-6}	5.53	3.82×10^{-10}
Keyword	KW-0805	Transcription regulation	2.16	1.84×10^{-6}	3.22	3.05×10^{-9}

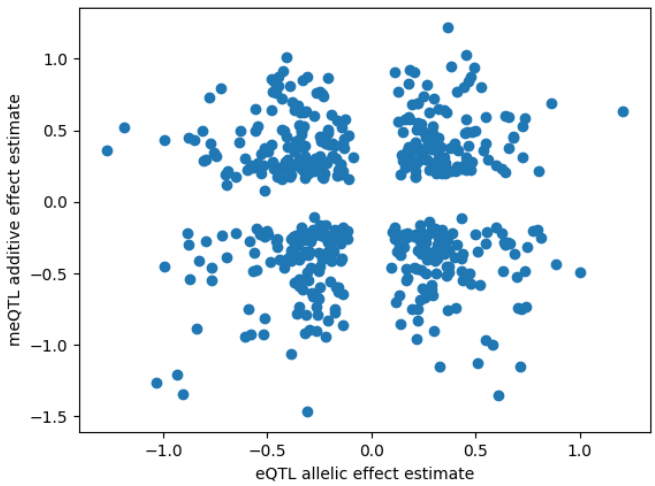
All known transcription factors

The 3,364 significant (discovery set; Bonferroni correction: $5x10^{-8}/512,601$) inter-chromosomal meQTLs we identify were also *locally*- acting eQTL for 835 unique Ensembl gene IDs in whole blood in GTEx (v6p) [13]. Of these, 599 are reported to be protein coding by Ensembl (GRCh37). 15.7% (94/599) of the protein coding genes are recognised transcription factors [30], and, of these, 77.7% (73/94) are C2H2-ZF proteins (Table 4.3). Of those identified as transcription factors, we find a non-replicated association between the additive meQTL effect estimate and the allelic effect eQTL effect estimate (Equation 4.2; p-value $9.97x10^{-4}$ and 0.22, set 1 and set 2, respectively) and no significant relationship with respect to dominance (Equation 4.3; p-value 0.66 and 0.93, set 1 and set 2, respectively). It is however worth noting that we are not surveying the full distribution of meQTL, or eQTL model p-values in this analysis and so these results should be interpreted with caution (Figure 4.3).

Figure 4.3: eQTL allelic effect estimates to meQTL effect estimate plots

All significant (discovery set; Bonferroni correction; <5x10^{-8}/512,601) inter-chromosomal meQTL that co-localised with a *locally*- acting eQTL in whole blood from GTEx (v6p) [13]. Limited to unique Ensembl Gene IDs, and one transcript per SNP.

A)



B)

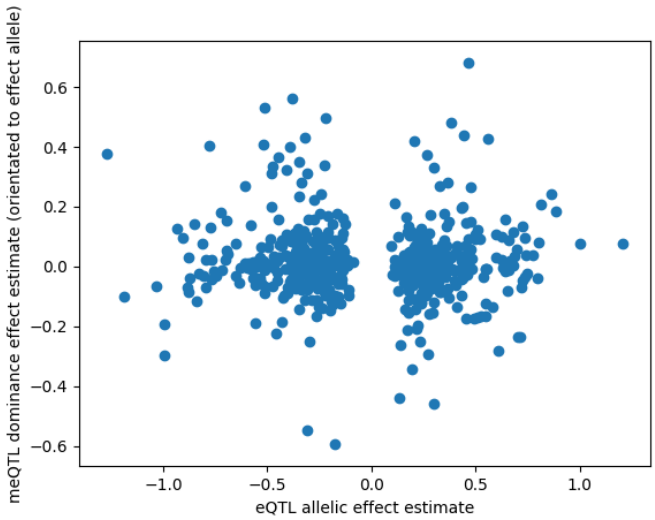


Table 4.3: Binding domains of the known transcription factors identified from significant inter-chromosomal meQTLs as significant locally- acting eQTL, protein-coding only

'Frequency: meQTL': the frequency of the DNA binding domain amongst those transcription factors identified for which the inter-chromosomal meQTL (discovery set) is a *locally*- acting eQTL in whole blood.

'Frequency: all known': the frequency of the DNA binding domain amongst all transcription factors identified by Lambert et al. [30].

DNA binding domain	Frequency meQTL (all known)	DNA binding domain	Frequency meQTL (all known)
C2H2	73 (759)	Forkhead	1 (49)
bHLH	4 (108)	GATA	1 (10)
Homeodomain	3 (228)	HSF	1 (8)
bZip	2 (54)	mTERF	1 (4)
IRF	2 (9)	Paired box	1 (9)
CG-1	1 (2)	T-box	1 (17)
DM	1 (7)	Unknown	1 (69)
FLYWCH	1 (1)	**Total**	**94 (1,686)**

In addition, of those proteins not identified as transcription factors, 18.9% (7/37) are identified as low specificity DNA binding proteins, and 16.2% (6/37) are identified as ssDNA/RNA binding.

It is noteworthy that, of the 599 protein coding genes that we identified as linked to inter-chromosomal meQTLs, 107 (17.9%) are confirmed as DNA or RNA binding by orthogonal methods, and 94 (15.7%) are recognised as bona fide transcription factors in the most comprehensive catalogue available to-date [30].

Interestingly, ZNF718 (ENSG00000250312) was not included in these results as it was incorrectly identified as a lincRNA in Ensembl GRCh37 [155] [accessed 25/06/2019], since updated in Ensembl GRCh38 [162] [accessed 24/10/2019] to protein coding. It is however, a C2H2-ZF protein and was classified as a transcription factor in the results of Lambert et al. [30].

Strikingly, these results represents 5.8% of all known or likely human transcription factors (95/1639), 9.7% (74/759) of all known human C2H2-ZF transcription factors, and 14.8% (53/357) of all known human Krüppel associated-box (KRAB) containing transcription factors [30]. C2H2 DNA binding domains are strongly over-represented, comprising 77.9% (74/95) of those transcription factors identified as associated with a *locally*- acting eQTL that co-localises with an inter-chromosomal meQTL (in the LD and CpG correlation filtered discovery set), as compared to 45.0% (759/1,686) of all known human transcription factors [30] (odds

ratio 4.30, Fisher's exact p-value 2.62×10^{-10}). As are KRAB containing transcription factors, comprising 55.8% (53/95), and 21.8% (357/1,639), respectively (odds ratio 4.53, Fisher's exact test p-value 3.59×10^{-12}).

Trans- acting hubs

As can be seen from Figure 4.2, there are clear *trans*- acting 'hubs' or 'hotspots' (vertical lines on the plot), where a single SNP (or group of closely located SNPs) influences the methylation level at multiple CpG sites. This recapitulates, albeit at greater resolution, a known meQTL result [12], and one that parallels expression data [169]. Taken as representative of C2H2-ZF containing proteins, the 74 proteins present in our network that were in the PFAM 'Zinc finger, C2H2 type' (PF00096) class (Table 4.3) were tagged by 76 independent (LD r^2 <0.2) SNPs. Of these 76 SNPs, 46% (35/76) were present on chromosome 19. This may contribute to the enrichment of *trans*- acting meQTLs found on chromosome 19, seen here (Figure 4.2) and previously [12].

Factor binding site CpG methylation under inter-chromosomal dominance genetic control

There are a number of theories that have been proposed to explain the molecular basis of dominance [170]: 1) dominant-negative mutations, 2) haploinsufficiency, and 3) gain-of-function mutations. Whilst it is sufficient for these mechanisms to be intra-locus to generate dominance effects, it is not necessary, and there is emerging interest in the genesis of dominance effects by regulatory networks and inter-locus interactions [170,171]. Classically, given a shared nuclear environment, *trans*- genetic effects are felt to imply a diffusible

mediator rather than direct genomic interaction, further supporting the hypothesis of inter-locus interaction. Transcription factors are the canonical *trans*- acting molecules and can affect multiple sites within the genome in both a sequence and methylation specific manner [140,141] and are of clear importance in cellular regulatory mechanisms.

Of the 3,364 significant (Bonferroni correction: p-value $<5x10^{-8}/512,601$; Table 4.1) inter-chromosomal SNP:CpG pairs (set 1) and 1,333 significant (Bonferroni correction: p-value $<5x10^{-8}/512,601$) pairs (set 2), 302 and 122 pairs, respectively, show a significant (Bonferroni correction: p-value <0.05/3,364, and <0.05/1,333 respectively) dominance component and 3,161, and 1,294 a significant (Bonferroni correction: p-value <0.05/3,364 and <0.05/1,333) additive component.

As I have demonstrated, C2H2-ZF transcription factors , especially KRAB domain containing transcription factors , as well as other transcription factors are implicated in the inter-chromosomal genetic control of DNA CpG methylation. Given this, I hypothesised that inter-chromosomal genetic control of DNA CpG methylation, especially dominance control, of CpG methylation at the Chip-Seq peaks of a given protein may implicate the protein itself, or its binding partners, as the cause of the variation in methylation.

With respect to Chip-Seq results (Codex [167]) and histone marks (Cistrome [166] Epigenome), replicated associations are uniformly associated with increasing dominance effects on CpGs within Chip-Seq peaks (Supplementary Table 11). This implies that, for CpG within the Chip-

Seq peak, the methylation status of the heterozygote is more similar to the more methylated homozygote (Figure 4.1). Replicated histone marks associated with increased dominance genetic control of CpGs within the locus (Bonferroni correction: discovery p-value <0.05/21; replication p-value <0.05/9) included: Acetylated H3, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H3K9K14ac, and H3K14ac. Replicated Chip-Seq peaks associated with increased dominance genetic control of CpGs within the locus are in Table 4.4. Interestingly, when absolute magnitude of additive effect was analysed as the dependent variable (rather than magnitude of dominance effect; i.e. Equation 4.5 rather than Equation 4.4), there was only one replicated result: an association between H3K9K14ac and absolute additive effect.

Finally, we used data from ENCODE to assess the seven genomic-segmentation states of the combined ChromHMM and Segway segmentation [163,164,168] and found strong replicated (Bonferroni correction: discovery p-value <0.05/14; replication p-value <0.05/4) evidence for an association of CpG location within predicted transcription start-sites and increased dominance control (discovery p-value $<2x10^{-16}$; replication p-value $<2x10^{-16}$), transcribed regions and reduced dominance control (discovery p-value $6.35x10^{-15}$; replication p-value $6.20x10^{-8}$), and weaker, but still clearly replicated evidence, for predicted repressed or low activity regions (discovery p-value $3.01x10^{-9}$; replication p-value $3.74x10^{-3}$) indicating a complex relationship between the inter-chromosomal meQTLs, their target sites, and predicted transcription profiles. The Chip-Seq peaks that we identify (Table 4.4), and the transcription factors we identify as acting between chromosomes do not overlap (match based on HGNC symbol). This implies either that there is modulation of the DNA CpG methylation within the region, for example through spreading of DNA CpG methylation from

the binding sites of the KRAB factors to encompass the binding sites of these transcription factors, or limitations due to lack of statistical power. Note that this result does not impact those Chip-Seq analyses the target of which is not a transcription factor.

Table 4.4: Replicated Chip-Seq peaks associated with increased dominance genetic control of CpGs within the locus.

Replicated Chip-Seq peaks associated with increased dominance genetic control of CpGs within the locus. Discovery p-value threshold: $<5.21 \times 10^{-4}$ (0.05/96); Replication p-value threshold: $< 1.39 \times 10^{-3}$ (0.05/36) (Bonferroni correction). Direction of effect was positive in all cases.

Antibody target (HGNC Symbol)	**Antibody target (HGNC Approved name)**	**Transcription Factor?**	**Comments** [30]
CBFB	core-binding factor subunit beta		CBFB is a co-factor: augments the binding of RUNX to DNA [172].
CDK7	cyclin dependent kinase 7		Serine/threonine kinase involved in cell cycle control and in RNA polymerase II-mediated RNA transcription. Phosphorylates, amongst other things, POLR2A [158] [accessed 24/10/2019].
EP300	E1A binding protein p300		EP300 is a co-factor. It has many motifs, presumably because many transcription factors recruit it.
ERG	ETS transcription factor ERG	Yes	
ETS1	ETS proto-oncogene 1, transcription factor	Yes	
HDAC1	histone deacetylase		HDAC1 is likely to be a co-factor.

MED21	mediator complex subunit 21		MED21 does not contact DNA in the structure of the yeast mediator complex (PDB: 5SVA)
MYB	MYB proto-oncogene, transcription factor	Yes	
MYH11	myosin heavy chain 11		Assayed in leukaemia cell lines harbouring inv(16) translocation (CBFB-MYH11). Therefore, it can be thought of as also targeting CBFB.
NOTCH1	notch receptor 1		A transmembrane protein that inhibits MEF2C activation [173]. NOTCH1 does not bind DNA in the structure PDB:4J2X, nor is it required for DNA binding by RBPJ.
PHF8	PHD finger protein 8		Histone lysine demethylase, with selectivity for mono- and di-methyl states [158] [accessed 24/10/2019].
POLR2A	RNA polymerase II subunit A		
RUNX1	RUNX family transcription factor 1	Yes	
RUNX3	RUNX family transcription factor 3	Yes	

TCF3	transcription factor 3	Yes
TCF12	transcription factor 12	Yes

Discussion

I identify many CpGs in Chip-Seq peaks that appear to be under inter-chromosomal dominance genetic control (Supplementary Table 11). In addition, I identify, using only meQTL and eQTL data, 5.8% of all known human transcription factors, 9.7% of all known human C2H2-ZF transcription factors, and 14.8% of all known human Krüppel associated-box (KRAB) containing transcription factors as having an eQTL that controls their abundance that also co-localises with an inter-chromosomal meQTL.

I demonstrate significant enrichment for C2H2-ZF and KRAB containing proteins in the set of inter-chromosomal meQTL associated *locally*- acting eQTL, when compared to a background of proteins associated with significant eQTL of whole blood (Table 4.2), as well as a significant enrichment of C2H2-ZF DNA binding domain and KRAB containing transcription factors amongst all known human transcription factors (odds ratio 4.30, Fisher's exact p-value 2.62×10^{-10}, and odds ratio 4.53, Fisher's exact p-value 3.59×10^{-12}, respectively).

KRAB zinc-finger proteins are encoded by an ancient group of genes, first emerging about 400 million years ago in a common ancestor of coelacanths, lungfish and tetrapods [174,175]. KRAB zinc-finger proteins are the largest family of transcriptional regulators in higher vertebrates [175], and humans encode 357 [30]. The proposed molecular functions of KRAB zinc-finger proteins include transcriptional repression of the promoters of RNA polymerases I, II, and III [176], RNA binding and splicing [176], and a complex 'domestication' of transposable elements [175]. Many, although not all [175], KRAB zinc-finger proteins interact with KAP1 to

form a repression complex containing the histone methyltransferase SETDB1 [177] the nucleosome remodelling and deacetylation (NuRD) complex [178], heterochromatin protein 1 (HP1) [179–182], and DNA methyltransferases [31]. Quenneville et al. [31] have previously demonstrated that, in embryonic stem cells, KAP1 mediates DNA CpG methylation, and that this methylation spreads over short distances (3 to 5kb; or further in other studies [32]) so as to involve nearby CpG islands. In addition, in a murine conditional gene regulation system, KRAB-mediated repression was seen to result in promoter DNA methylation and irreversible gene silencing when allowed to occur in early mouse embryogenesis [183]. However, the same phenomenon was not observed in differentiated cells, suggesting a temporal context to the repression. That is, in differentiated cells, the KRAB/KAP1-induced heterochromatin formation did not lead to corresponding DNA CpG methylation changes. Our results represent a potential extension of this framework to haematopoietic stem cells and their progeny. Indeed, the role of KRAB zinc-finger proteins in mitophagy during erythrogenesis has recently been established [184]. On top of this temporal context, previous work has demonstrated that repression by KRAB factors also demonstrates specificity with respect to genomic context [185]. The results presented here demonstrate one possible solution to the problem of identifying the native genomic targets, and hence contexts, of endogenous KRAB domain containing transcription factors.

All the transcription factors identified as co-localising with an inter-chromosomal meQTLs here, are thought to bind as a monomer or homomultimers to DNA [30]. However, the binding motif is unknown in 27.4% (26/95) [30]. The results presented here clearly provide a potential method for identifying the CpG targets of these factors. A recent work identifying the

impact of DNA CpG methylation on binding specificities of human transcription factors examined whether the motifs they identified when DNA CpG methylation was considered was similar or dissimilar to that identified by previous studies. 28.5% of the motifs identified that were different from that identified in preceding studies were C2H2-ZF proteins, compared to only 8.5% of those that were similar [141]. The cause of this discrepancy was commonly because the transcription factors binding specificity had not previously been determined, or whose preference for methylated cytosine had not previously been appreciated [141]. Of the 759 known C2H2-ZF transcription factors [30], about 100 were analysed by Yin et al. [141], and of these approximately one third were shown to preferentially bind some methylated sequences over the corresponding unmethylated sequence. However, with respect to binding affinity, a methylated cytosine may be more similar to a thiamine than to an unmethylated cytosine in its binding motif [186], and so this situation is more complex than simply comparing methylated to unmethylated, but otherwise identical, sequences.

A limitation of this study is that we were unable to distinguish between different types of cytosine methylation, e.g. 5-methyl-cytosine, 5-hydroxy-methyl-cytosine, 5-formyl-cytosine, and 5-carboxy-cytosine, the differences in which may impact transcription factor binding and regulatory feature response [187]. Therefore, meQTL that are specific to a particular methylation subtype may be missed. In addition, as with any study, we are faced with the impossibility of excluding all confounders, e.g. residual cell-type composition bias. However, only 12% of KRAB-containing C2H2-ZFs are thought to be tissue specific [30]. This, together

with the abundant nature of the meQTLs identified, make this explanation of the results unlikely.

C2H2-ZF proteins contribute the greatest amount of motif diversity to the pool of motifs recognised by human transcription factors [30]. It has previously been shown that KRAB zinc-finger protein binding correlates with H3K9me3 at their target loci [31,174]. We demonstrate that, as with all other replicated Chip-Seq peaks (Supplementary Table 11), H3K9me3 was associated with increased dominance genetic control of DNA CpG methylation.

It is estimated that approximately 93% of common disease-associated genetic variation is non-coding [11] and it has been shown that disease-associated genetic variation can have *distantly*- acting (>5Mb SNP to CpG distance, or located on different chromosomes) effects on DNA CpG methylation [15]. Indeed, it was found that approximately one third of trait-associated variants affected methylation at distant sites [15]. Amongst the inter-chromosomal meQTLs identified in our study, we have shown strong enrichment for C2H2-ZF proteins, especially those containing KRAB domains. Given the genome-wide coverage of both the SNP and CpG sites in this study, it enables the possibility of using the results as a dictionary to look up both the likely local and distant effects of genetic variation found to be disease associated. That is, one could take the SNPs identified as disease associated from a GWA study and attempt to identify which other genomic locations are affected by their variation by mapping the location of the CpGs affected by their variation, in an analysis similar to that mentioned in Chapter 3 with respect to the effects of pQTL on methylation.

Conclusions

I have identified 5.8% of all known human transcription factors – including 9.7% of all known human C2H2-ZF protein transcription factors, and 14.8% of all known human Krüppel associated-box (KRAB) containing transcription factors – as affected by a *locally*- acting eQTL that is also an inter-chromosomal meQTL. I also highlight the abundance of non-additive genetic effects driving variation of DNA CpG methylation at distant CpGs. I review the enrichment of these loci for the binding of many protein complexes and find this to be common.

Based on genomic location, it has previously been proposed that *distantly*- acting meQTL co-localise with genomic regions containing zinc-finger containing transcription factors [21]. I extend this result and find strong enrichment for variation in the expression of C2H2-ZF proteins (using *locally*- acting eQTL data), especially those containing a Krüppel associated box (KRAB) domain. In addition, I go on to highlight numerous protein complex binding sites within which DNA CpG methylation is enriched for inter-chromosomal dominance genetic control.

In this chapter I have illuminated a small part of the complex regulatory landscape of the cell. This work has important implications for the unpicking and interpretation of non-coding GWA variants in both health and disease. As much of GWA disease associated variation is in

non-coding regions, understanding the regulatory networks of the cell is of critical importance in the post-GWA study era.

Just as no man is an island, neither is a CpG, rather it is a part of an integrated (epi-)genomic whole.

5. Conclusion

Summary of main findings

This thesis includes three studies of intermediate traits, and examines their utility in illuminating the pathways between genetic variation and phenotype. I undertook the assessment of both plasma and cellular protein abundance, as well as DNA CpG methylation.

In Chapter 2, I conducted the genome-wide association of 249 proteins, as measured in plasma, using an antibody-based assay in two European populations. I used the results of this to perform proteome-by-phenome MR and demonstrated 509 putative causal links between various proteins and outcome diseases and traits, including such links to cardiovascular disease and schizophrenia.

In Chapter 3, I demonstrated the feasibility of examining the cellular proteome of peripheral blood mononuclear cells and presented an exploratory analysis of the results of the first 251 samples. I compared these results with those found for RNA expression in whole blood from the GTEx (v6p) project [13]. This chapter represents the deepest assessment of the proteome of human peripheral blood mononuclear cells at a population-scale to date. We find evidence for a causal contribution of both members (PADI2 and PADI4) of the 'histone H3-R26 citrullination' pathway (GO:0036413) in rheumatoid arthritis, as well as demonstrate a putative method for assessing pleiotropic effects of any given SNP using meQTL data.

In Chapter 4, I unpacked the molecular mechanisms underlying inter-chromosomal meQTL by combining the results generated here with those of the GTEx (v6p) project [13]. I identified 95 *locally*- acting eQTL for transcription factors that are also inter-chromosomal meQTL. Among these I demonstrated a significant enrichment for C2H2-ZF and Krüppel associated-box (KRAB) containing transcription factors. In addition, I demonstrated an enrichment for C2H2-ZF and KRAB containing transcription factors amongst eQTL that co-localise with an inter-chromosomal meQTL amongst all significant whole blood *locally*- acting eQTL in GTEx (v6p) [13]. This suggests that C2H2-ZF and KRAB containing transcription factors may be the molecular intermediates mediating the effects of inter-chromosomal meQTL. In addition, I find that CpG affected by an inter-chromosomal meQTL may be found in those parts of the DNA that interact with many different protein complexes, suggesting that *trans*- acting effects on methylation may be widespread and not just limited to the direct effects of transcription factors, but of their binding partners, and the entire regulatory network.

Limitations

An easy dichotomy to draw is between theoretical and practical limitations encountered.

Theoretical

First and foremost, there is the issue of statistical power, in particular with respect to the two proteomics studies presented here. However, sample size in all cases was governed by external constraints. An obvious solution to this is meta-analysis and my consortia contributions are mentioned in the preface to this thesis. Thankfully, given the initial promising results of the peripheral blood mononuclear cell proteomics project, when

complete the sample size for this study will be approximately 3.4 times larger than that discussed here. Compounding issues of statistical power, in the context of multiple correlated outcomes a Bonferroni correction is likely to be extremely cautious and will clearly have downstream consequences for those SNPs selected for follow-up. In order to address this, at various points throughout this thesis, a more permissive FDR threshold has been used.

Throughout this thesis, the 1,000 Genomes [92,93] dataset was used as an LD reference. Due to the reduced number of SNPs present in this data when compared to the HRC [104] imputed data used for genome-wide association, there will have been a consequent loss of resolution when picking lead-SNPs for downstream analyses.

Fundamental to the legitimacy of the results of MR are the following assumptions: 1) that the SNP is associated with the exposure of interest, 2) that the SNP is independent of any confounders, and 3) that the SNP does not influence the outcome of interest, except via the exposure variable [26]. It is also assumed that the relationship between the exposure and outcome is linear [26]. Assumption 3 could be stated another way, that is, that the genetic variant does not exhibit horizontal pleiotropy. Assumptions 2 and 3 are clearly untestable given the impossibility of measuring all possible confounders and pathways between instrumental variable and outcome. However, *a priori* knowledge of the biology of the system can be leveraged to, in part, address concerns about their validity. In addition, as alluded to in Chapter 3, by increasing the depth to which the proteome is assessed, it is possible to enact a more thorough assessment of the pleiotropic effects of any given SNP by

virtue of measuring more of the potential intermediate variables between SNP and outcome trait. For example, in this context, these intermediate traits could be all the proteins encoded within 1Mb of the SNP.

In addition, while a significant MR result implies that the exposure affects the outcome, it does not provide specific information as to when the protein exerted its effect, nor the relevant tissue within which this modulation occurred. Both of these influence the practicality of using a particular protein as a therapeutic target.

Finally, at various points in this thesis an appeal to *a priori* biological knowledge has been used to make inference regarding direction of causation. That is, that the proximity of a genetic variant altering the abundance of an RNA or a protein to the gene encoding that RNA / protein implies a direct causal link between the variant and RNA / protein abundance. However, this is not absolute, and reverse causation (e.g. SNP -> disease outcome -> protein abundance) cannot be excluded. Methods of causal discovery (that is to attempt to ascertain the underlying causal graph from the data) do exist in the context of Mendelian Randomisation, for example bidirectional (reciprocal) MR [188], and MR Steiger [189], and may be useful in future work.

Practical

Given the number of the entities measured by '-omic' technologies the computational challenges in their analysis are substantial. This poses problems both in the processing of

these data and the subsequent storage of the results. For example, despite compression of the output data, and restriction of the number of significant digits stored, we were only able to store SNP:CpG results with a model p-value of $<1x10^{-3}$ for the meQTL analyses. In filtering the results for storage in this manner, it does unfortunately limit the use of these data in analyses that require full summary statistics. However, given some assumptions about the distribution of these data, it may be possible to impute the missing results in the $1x10^{-3}$ to 1 range.

As discussed in Chapter 4, the binding specificities of antibody-based assays (such as the Olink assay used in Chapter 2) are not perfect and are vulnerable to any variation that causes a conformational change in the protein that affects the epitope of the antibody. An analogous issue exists with the mass-spectrometry results in that non-synonymous changes to the protein will potentially change its mass, elution profile, and digestion fragments. However, for directly measured – or well imputed – non-synonymous variants in the mass-spectrometry data-set, this is a surmountable problem. In an individual harbouring a non-synonymous variant, the mutant protein is likely to have been directly measured and recorded in the raw mass-spectrometry data. It is theoretically possible to include all non-canonical protein sequences in the peptide / protein search of the output data. However, the computational and multiple testing burdens of such and analysis should not be underestimated. Finally, unknown proteins cannot be quantified using the current method of searching that matches measured spectra to those predicted by an *in-silico* digest of all known proteins. This reinforces the necessity to analyse features directly.

Potential Future Directions

Proteomics

As the peripheral blood mononuclear cell work is an exploratory analysis and only approximately 30% of the predicted final sample size, the entire chapter is forward-looking. There are many exciting future possibilities with this data set, as summarised in Chapter 4 and briefly recapitulated here: analysis of the correlation structure of the proteins and outlier removal, the investigation of alternative imputation strategies, and the inclusion of further cell-type specific markers in the regression model (for example CD3, CD19, and CD14 to tag T-lymphocytes, B-lymphocytes, and monocytes). Formal heterogeneity testing, such as HEIDI [27], as well as the investigation of novel methods such as that involving assessment of the meQTL over a locus, should be considered. It is worth noting that, as the breadth of coverage increases, the opportunity to directly assess pleiotropic effects within a locus will increase. That is, one can directly assess the association between a given genetic variant and those proteins that have been measured, as was the case for three of the five human PADI enzymes in Chapter 3.

As mentioned above, GWA of the high-performance liquid-chromatography mass-spectrometry features themselves is likely to be a profitable endeavour, as is a standard search of these data with additional allowable variable protein modifications (i.e. post-translational modifications) and a non-synonymous variant aware search. Sample preparation for mass-spectrometry is designed to enrich the features present for peptide. This means that direct assessment of the genetic control of the features themselves enables the assessment of the genetic control of all the peptides detected in the sample,

irrespective of whether or not they have been identified. Formal identification of a feature can be undertaken retrospectively and may require an additional mass-spectrometry run with selection of the feature for fragmentation if it has not been fragmented in any of the samples or the library. If fragmentation has already been performed and the feature remains unidentified, then identification may be possible with the inclusion of additional modifications or protein sequences (e.g. non-synonymous mutations) in the database search; if identification remains elusive, then *de novo* sequencing may provide an identification. *De novo* sequencing of otherwise unidentified features, together with their genetic associations provides a means of identifying novel peptides, as well as the genetic loci that regulate their production.

In addition to enabling the identification of outliers, between protein correlations, independent of the genetics, would allow for data-driven clustering of the proteins themselves, these data-driven groupings are likely to represent the true underlying biological pathways [4]. Assessment of enrichment of such data-driven clusters with known biological pathways would provide a simple method of experimental, from myriad orthogonal sources, validation for such groupings.

DNA CpG methylomics

With respect to this body of work, there are many potentially fruitful future possibilities to pursue, including: 1) an assessment of commonality between those CpG sites under the control of SNPs associated with specific transcription factors, especially C2H2 and KRAB

containing transcription factors and, conversely, 2) is there any commonality in the molecules that mediate specific subsets of SNP:CpG pairs, for example, is there an enrichment of a specific protein family in those *trans*- acting meQTL for which the CpG is located in an enhancer element?, 3) is there an enrichment of a particular RNA species, i.e. the non-protein coding transcripts, identified as inter-chromosomal meQTLs?, and 4) in an analogous manner to inter-chromosomal meQTL, further investigation into the molecular mechanisms underpinning *distantly*- and *locally*- acting meQTL may be warranted.

C2H2-ZF, including KRAB domain containing, transcription factors were enriched amongst sites associated with a *locally*- acting eQTL and inter-chromosomal meQTL when compared to all *locally*-acting eQTL in whole blood. One potential explanation for this is that the genetic variation affects the expression of the transcription factor locally, and the transcription factor then subsequently changes the transcriptional profile and methylation status at distant sites. To state this hypothesis explicitly, I suggest that the CpG location of the *trans*- acting meQTL indicate the genomic location(s) affected by the transcription factor encoded locally to the SNP, noting that this may be through direct or indirect mechanisms. Further support for this hypothesis could be obtained from existent eQTL data, that is, it could be assessed as to whether the *trans*- acting meQTL are also *trans*- acting eQTL for a transcript in the vicinity of the distant CpG. As noted previously, eQTL studies are often small and statistical power limited to detect *trans*- acting associations, however, this analysis would be more targeted than a simple genome-wide scan, at least in part mitigating this issue. Other potential confirmatory experiments could include targeted mutation of the

KRAB domain of a given transcription factor in a human cell line, followed by assessment of the methylation status at the relevant CpG sites (as indicated by the meQTL data).

Final words

This thesis has examined, in depth, the genetic effects controlling two intermediate traits – protein and DNA CpG methylation – at unprecedented scale in both arenas. I have taken these results and projected them onto, amongst other things, disease risk. I have, with moderate success, identified potential novel therapeutic targets and described previously unknown biology. I have demonstrated the benefit of using combined '-omics' data, and hope to contribute a comprehensive assessment of the cellular proteome of peripheral blood mononuclear cells in the near future.

6. Appendices

Appendix 1: Materials relating to Chapter 2

Supplementary Table 1. List of pQTLs (linkage disequilibrium clumped): indep_pqtl.tsv.

List of lead-SNPs for each protein following linkage disequilibrium (LD) clumping, together with replication information. Biallelic variants within ±5Mb and r^2 >0.2 to the lead variant (smallest p-value at the locus) were clumped together. European populations in 1,000 Genomes[92,93] were used as the LD reference.

Columns are: 'hgnc_symbol': HUGO gene naming consortium symbol of the exposure (protein); 'snpid': 'chr'_'pos'; 'rsid': rsID; 'chr': chromosome (GRCh37) of the SNP; 'pos': position (GRCh37) of the SNP; 'a1': effect allele; 'a0': other allele; 'n_pri': number of individuals in the primary cohort (CROATIA-Vis); 'freq1_pri': frequency of the effect allele is the primary cohort (CROATIA-Vis); 'beta1_pri': beta estimate of the effect allele in the primary cohort (CROATIA-Vis); 'se_pri': standard error of 'beta1_pri' in the primary cohort (CROATIA-Vis); 'p_pri': p-value of 'beta1_pri' and 'se_pri'; 'info_pri': SNPTEST (v2) info of the imputation in the primary cohort (CROATIA-Vis); 'r2_pri': coefficient of determination of the regression in the primary cohort (CROATIA-Vis); 'n_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'freq1_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'beta1_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'se_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'p_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'info_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'r2_sec': as for the

primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'uniprot_swissprot': UniProtID of the exposure (protein), see http://www.uniprot.org/; 'ensembl_gene_id': Ensembl gene ID (GRCh37; see http://grch37.ensembl.org/index.html) of the gene-of-origin of the protein; 'chromosome_name': chromosome (GRCh37) of the gene of the protein, as per Ensembl GRCh37; 'start_position': start position (GRCh37) of the gene of the protein, as per Ensembl GRCh37; 'end_position': end position (GRCh37) of the gene of the protein, as per Ensembl GRCh37; 'description': HUGO gene naming consortium description of the exposure (protein); 'replicated_pqtl': is the lead-SNP of the cluster (as identified in the primary cohort) replicated in the secondary cohort (Bonferroni correction for multiple testing. TRUE if it is; FALSE if not); 'within_gene_plus_flank_tol': is the SNP within the gene-of-origin of the protein +/- 150kb (TRUE is it is; FALSE if not).

Supplementary Table 2. Comparison of the lead-SNPs identified here and eQTL: 'pQTL_eQTL.tsv'.

eQTL data derived from 'Whole blood' from GTEx[13] (v7). Bonferroni correction 0.05/54.

Columns are 'hgnc_symbol': the HGNC symbol corresponding to the UniProtID; 'rsid': rsID of the SNP; 'chr': chromosome of the SNP, GRCh37; 'pos': position of the SNP, GRCh37; 'a1': the effect allele; 'a0': the other allele; 'uniprot': UniProtID of the protein; 'n_protein_pri': number of individuals in the primary protein cohort (CROATIA-Vis); 'freq1_protein_pri': frequency of the effect allele in the primary protein cohort (CROATIA-Vis); 'beta1_protein_pri': effect-size estimate in the primary protein cohort (CROATIA-Vis); 'se_protein_pri': standard error of 'beta1_protein_pri'; 'p_protein_pri': p-value of 'beta1_protein_pri' and 'se_protein_pri'; 'info_protein_pri': SNPTEST (v2) imputation info score in the primary protein cohort (CROATIA-Vis); 'n_protein_sec': as for the primary

cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘freq1_protein_sec’: as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘beta1_protein_sec’: as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘se_protein_sec’: as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘p_protein_sec’: as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘info_protein_sec’: as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); ‘ensembl_gene_id’: Ensembl gene ID corresponding to the protein; ‘pval_nominal_gtex’: nominal p-value in GTEx (v7) whole blood; ‘slope_gtex’: effect-size estimate in GTEx (v7) whole blood; ‘slope_se_gtex’: standard error of ‘slope_gtex’ in GTEx (v7) whole blood; ‘pval_nominal_threshold_gtex’: nominal p-value threshold for calling a variant-gene pair significant for the gene in GTEx (v7) whole blood; ‘min_pval_nominal_gtex’: smallest nominal p-value for the gene in GTEx (v7) whole blood; ‘pval_beta’: beta-approximated permutation p-value for the gene in GTEx (v7) whole blood.

Supplementary Table 3. Comparison of the lead-SNPs identified here and those identified using an orthogonal, aptamer-based assay: ‘ld_list_olink_sun_with_ld.tsv’.

Aptamer-based assay results are those of Sun et al.[14].

Columns are ‘hgnc_symbol’: the HGNC symbol corresponding to the UniProtID; ‘exposure’: the UniProtID of the protein; ‘rsid_olink’: the rsID of the lead-SNP from this study; ‘chr_olink’: the chromosome, GRCh37, of the lead-SNP from this study; ‘pos_olink’: the position, GRCh37, of the lead-SNP from this study; ‘a1_olink’: allele 1 of the lead-SNP from this study; ‘a0_olink’: allele 0 of the lead-SNP from this study; ‘rsid_sun’: the rsID of the lead-SNP from Sun et al.; ‘chr_sun’: the chromosome, GRCh37, of the lead-SNP from Sun et

al.; ‘pos_sun’: the position, GRCh37, of the lead-SNP from Sun et al.; ‘a1_sun’: allele 1 of the lead-SNP from Sun et al.; ‘a0_sun’: allele 0 of the lead-SNP from Sun et al.; ‘ld_r2’: the linkage disequilibrium (r^2) of the two SNPs, as measured in the European individuals from 1,000 Genomes (Methods).

Supplementary Table 4. Additional studies identified using Phenoscanner: additional_studies.tsv.

Table of the additional studies (and outcome traits) identified through Phenoscanner[67,68]. Note that ‘Coronary artery disease’ was included from van der Harst et al.[61] both with and without the inclusion of data from UK Biobank.

Columns are ‘Outcome’: trait under study; ‘PMID’: Pubmed ID of the study; ‘First author’: First author the publication; ‘Year’: year of publication of the study; ‘Paper title’: title of the study.

Supplementary Table 5. Mendelian Randomisation results from GeneAtlas: df_ukbb_heidi.tsv.

Table of the all significant (FDR <0.05) Mendelian Randomisation (MR) results using data from GeneAtlas[29] together with their HEIDI[27] test statistic results. pQTL for both cohorts are included, however, in order to avoid a ‘winner’s curse’, MR and HEIDI were conducted using data from the secondary protein cohort (ORCADES).

Columns are 'hgnc_symbol': HUGO Gene Nomenclature Committee symbol of the exposure protein; 'outcome_description': description of the UK biobank outcome from GeneAtlas; 'rsid': rsID; 'snpid': 'chr'_'pos'; 'chr': chromosome (GRCh37); 'pos': position (GRCh37); 'a1': effect allele; 'a0': other allele; 'exposure': UniProtID of the protein;

'ensembl_gene_id': Ensembl (GRCh37) gene ID of the exposure protein; 'n_exposure_pri': number of individuals in the primary protein cohort (CROATIA-Vis); 'freq1_exposure_pri': frequency of the effect allele in the primary protein cohort (CROATIA-Vis); 'beta1_exposure_pri': regression coefficient (per additional effect allele) in the primary protein cohort (CROATIA-Vis); 'se_exposure_pri': standard error of 'beta1_exposure_pri'; 'p_exposure_pri': p-value of 'beta1_exposure_pri' and 'se_exposure_pri'; 'info_exposure_pri': SNPTEST (v2) imputation info score in the primary protein cohort (CROATIA-Vis); 'n_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'freq1_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'beta1_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'se_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'p_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'info_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'outcome': outcome code of the UK biobank outcome from GeneAtlas; 'beta1_outcome': beta of the effect allele on the outcome in GeneAtlas; 'se_outcome': standard error of 'beta1_outcome'; 'p_outcome': p-value corresponding to 'beta1_outcome' and 'se_outcome'; 'info_outcome': imputation info score in UK Biobank; 'freq1_outcome': frequency of the effect allele in UK Biobank; 'beta_mr_delta_sec': beta value using the delta MR method (using up to second order partial derivatives; See the appendix of Lynch and Walsh for further information) using estimates from the secondary cohort; 'se_mr_delta_sec': standard error of 'beta_mr_delta_sec' using the delta MR method (using up to first order partial derivatives; See the appendix of Lynch and Walsh for further information) using estimates from the secondary cohort; 'p_mr_delta_sec': p-

value corresponding to 'beta_mr_delta_sec' and 'se_mr_delta_sec'; 'fdr_sig_mr_delta_sec': significance of 'p_mr_delta_sec' at a False Discovery Rate (FDR) of <5%. True / False; 'p_HEIDI': p-value of the HEIDI statistic; 'nsnp_HEIDI': the number of SNPs used in the calculation of the HEIDI statistic.

Supplementary Table 6. Mendelian Randomisation results from studies identified using Phenoscanner: df_phenoscanner.tsv.

Table of all Mendelian Randomisation results using data acquired through Phenoscanner[67,68]. pQTL for both cohorts are included, however, in order to avoid a 'winner's curse', MR was conducted using data from the secondary protein cohort.

Columns are 'hgnc_symbol': HUGO Gene Nomenclature Committee symbol of the exposure protein; 'trait': outcome trait description; 'snp': chr'chr':'pos'; 'rsid': rsID; 'chr': chromosome (GRCh37); 'pos': position (GRCh37); 'a1': effect allele; 'a0': other allele; 'exposure': UniProtID of the protein; 'n_exposure_pri': number of individuals in the primary protein cohort (CROATIA-Vis); 'freq1_exposure_pri': frequency of the effect allele in the primary protein cohort (CROATIA-Vis); 'beta1_exposure_pri': regression coefficient (per additional effect allele) in the primary protein cohort (CROATIA-Vis); 'se_exposure_pri': standard error of 'beta1_exposure_pri'; 'p_exposure_pri': p-value of 'beta1_exposure_pri' and 'se_exposure_pri'; 'info_exposure_pri': SNPTEST (v2) imputation info score in the primary protein cohort; 'n_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'freq1_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'beta1_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'se_exposure_sec':

as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'p_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'info_exposure_sec': as for the primary cohort (CROATIA-Vis) but in the secondary cohort (ORCADES); 'ensembl_gene_id': Ensembl (GRCh37) gene ID of the exposure protein; 'study': name of the consortium/lead author of the outcome study; 'pmid': PubMed ID of the outcome study; 'ancestry': ancestry of the population within which the outcome was measured; 'year': the year the outcome study was published; 'beta1_outcome': regression coefficient (per additional effect allele) in the outcome study; 'se_outcome': standard error of 'beta1_outcome'; 'p_outcome': p-value of 'beta1_outcome' and 'se_outcome'; 'n_outcome': number of individuals in the outcome study; 'n_cases_outcome': number of cases in the outcome study; 'n_controls_outcome': number of controls in the outcome study; 'n_studies_meta_outcome': if a meta-analysis, number of studies included; 'units_outcome': units of analysis in the outcome study (IVNT stands for inverse normal rank transformed phenotype); 'dataset': Phenoscanner dataset ID; 'beta1_outcome_flipped': has the sign of 'beta1_outcome' been inverted from that provided by Phenoscanner due to calling of the effect vs. non-effect allele? True / False; 'beta_mr_delta_sec': beta value using the delta MR method (using up to second order partial derivatives; See the appendix of Lynch and Walsh for further information) using estimates from the secondary cohort; 'se_mr_delta_sec': standard error of 'beta_mr_delta_sec' using the delta MR method (using up to first order partial derivatives; See the appendix of Lynch and Walsh for further information) using estimates from the secondary cohort; 'p_mr_delta_sec': p-value corresponding to 'beta_mr_delta_sec' and 'se_mr_delta_sec'; 'fdr_sig_mr_delta_sec': significance of 'p_mr_delta_sec' at a False Discovery Rate (FDR) of <5% (True / False).

Supplementary Table 7. ChEMBL results: chembl_matches.tsv.

Compounds targeting the mediators listed in Supplementary Table 5. Columns are ‘uniprot’: UniProtID; ‘gene_symbol’: Gene Symbol; ‘target_chembl_id’: CHEMBL ID for this protein ; ‘compound_id’: CHEMBL compound ID; ‘max_phase’: CHEMBL-reported maximum phase of drug development for this compound; ‘drug_synonyms’: drug names; ‘indication_class’: CHEMBL-reported indication for this compound.

Supplementary Table 8. Key of Figure 2a: df_phenoscanner_outcomes.tsv.

Key for the abbreviations used in Figure 2a.

Columns are ‘Abbreviation’ and ‘Outcome Description’.

Supplementary Table 9: Key of Figure 2b: df_ukbb_outcomes.tsv.

Key for the abbreviations used in Figure 2b.

Columns are ‘Abbreviation’ and ‘Outcome Description’.

Appendix 2: Materials relating to Chapter 3

Protocol: preparation of cell lysate.

1. Collect samples from LN2, allow to thaw on ice.
2. Label 15 ml falcon tubes.
3. Prep Cat 2 room
    a. Get from outside: Ice, PBS, labelled 0.5ml centrifuge tubes, lysis buffer, beaker for tips in hood, boxes of filter tips: 1 x 200ul, 2 x 1000ul, 10ml stripettes, 15ml falcon tubes, dry ice.
    b. Prepare hood.
4. Add 9ml Phosphate Buffered Saline (PBS) to each 15ml falcon tube.
5. Ensure cells adequately suspended.
6. Transfer cells to labelled 15ml falcon.
7. Repeat steps 5-6 for each sample in batch.
8. Wash cells x2
    a. Spin sample 400g for 10min at 4degC
        i. First Batch: 1. A [__] B [__] 2. A [__] B [__]
        ii. Second Batch: 1. A [__] B [__] 2. A [__] B [__]
    b. Remove supernatant.
    c. Resuspend pellet in 10ml PBS.

9. Spin sample 400g for 10min at 4degC.
    a. First Batch: 3. A [__] B [__]
    b. Second Batch: 3. A [__] B [__]
10. Remove supernatant.
11. Resuspend pellet in 40ul of ‘lysis buffer’ (6M Guanidine HCl + 100mM Tris pH8.5).
12. Transfer to 0.5ml centrifuge tubes.
13. Place the samples on dry ice.
14. Prepare to exit the Cat 2 room.
15. Thaw samples.
16. Sonicate samples for 10sec 5um with Soniprep 150.
17. Place samples back on dry ice.

Protocol: Digest, day 1

1. Add 18ul H_2O to wells A1 : D12 of ‘protein assay plate’ (‘A’).
2. Thaw samples.
3. Spin samples to bottom of 0.5ml centrifuge tubes.
4. From the 0.5ml centrifuge tubes put 2ul of sample onto ‘protein assay plate’ (‘A’) and return 0.5ml centrifuge tubes to dry ice.
5. Perform Protein Assay.

a. Per batch: Make reagent A and B mix (20ml reagent A, 400ul reagent B).

b. Add 180ul of reagent A and B mix to all wells of 'protein assay plate' ('A).

c. Incubate for 30mins.

6. Calculate the volume of sample and the volume of lysis buffer (6M Guanidine HCl + 100mM Tris pH8.5) required to standardise amount of protein per well to 15ug (using spreadsheet template to create a manifest).

7. Thaw samples on ice.

8. BETWEEN BATCH REPEATS

   a. Transfer specified volume, as directed by the manifest, of lysis buffer onto the 'digest plate' ('D') **of subsequent digests.**

   b. Transfer specified volume, as directed by the manifest, of sample onto the 'digest plate' ('D') **of subsequent digests**.

   c. Place any digest plates that are not for immediate processing in the freezer (-20degC).

9. WITHIN BATCH REPEAT: Transfer the specified volume, as per the manifest, of lysis buffer of the sample to be repeated into D10 of the 'digest plate' ('D').

10. STANDARD SAMPLE: Transfer specified volume, as per the manifest, of lysis buffer onto the 'digest plate' ('D').

11. WITHIN BATCH REPEAT: Transfer specified volume, as per the manifest, of the sample to be repeated into D10 of the 'digest plate' ('D')

12. STANDARD SAMPLE: Transfer specified volume, as per the manifest, of sample onto the ‘digest plate’ (‘D’).
13. Transfer the remaining sample into to the ‘lysate storage plate’ (‘L’) and place in the freezer (-20degC).
14. Add 1ul of 100mM TCEP to each well of ‘digest plate’ (‘D’).
15. Add 1ul of 200mM CAA to each well of ‘digest plate’ (‘D’).
16. Place lid on ‘digest plate’ (‘D’).
17. Boil ‘digest plate’ (‘D’) (95degC for 5min).
18. Remove from block and allow to cool for 5 min.
19. Spin down condensate.
20. Add 20ul LysC working reagent to each well.
    a. LysC working reagent, per batch: resuspend 1 ampule (20ug) LysC in + 1333ul pre-prepared LysC buffer (100mM Tris pH 8.5).
21. Check pH [__]
    a. Ensure pH8 to 8.5.
22. Incubate at 37degC overnight (16 hours).

Protocol: Digest, day 2

1. Spin down condensate.

2. Make up trypsin working reagent.
    a. Per batch: 4ml Trypsin buffer (50mM Tris pH8.5) + 7.5ul Trypsin stock (1 ug/ul Trypsin in 0.1% TFA).
3. Add 80ul of trypsin working reagent to each well of ‘digest plate’ (‘D’).
4. Check pH [__]
    a. Ensure pH 8 to 8.5.
5. Incubate for 4 hours at 37degC.
6. Prepare C18 stage tips (with two pieces of C18 per tip).
7. Spin down condensate.
8. Add 15ul of methanol to each tip.
9. Spin tips: 300g, 2 min, 22degC.
10. Add 50ul of 0.1% TFA to each tip.
11. Spin tips: 300-500g, 5 min, 22degC.
12. Add 16ul of 10% TFA to each well of ‘digest plate’ (‘D’).
13. Check pH [__]
    a. Ensure <=3.
14. Transfer 120ul of samples from ‘digest plate’ (‘D’) to stage tips.
15. Spin tips: 500-900g, 5 min, 22degC.
    a. Repeat as necessary if liquid not fully through tip.

16. Add 50ul 0.1% TFA to tips.
17. Spin tips: 500-900g, 5min, 22degC.
    a. Repeat as necessary if liquid not fully through tip.
18. Place stage tips into adapter above ‘Elution plate’ (‘E’).
19. Add 40ul of 80% acetonitrile (ACN) + 0.1% TFA to tips.
20. Spin tips: 200-500g, 5min, 22degC.
    a. Repeat as necessary if liquid not fully through tip.
21. Put lid on plate for transfer to centrifuge.
22. Dry ‘Elution plate’ (‘E’) in vacuum centrifuge: 20min, 30degC, high-volatile (without lid).
    a. Repeat if not fully dry at the end of centrifugation.
23. Resuspend samples in 14ul of MS grade water.
    a. Wait 30min for adequate resuspension.
24. Transfer 5ul of each well of ‘Elution plate’ (‘E’) to ‘Peptide Assay plate’ (‘A’).
25. Add 1ul of 1% TFA to each well of ‘Elution plate’ (‘E’).
26. Check pH ~3.
27. Store peptide assay plate for later processing.
28. Thaw 20ul aliquot of pooled standard.
29. Spin pooled standard to bottom of tube.

30. Add pooled standard to D12 of elution plate.

31. Spin samples to bottom of ‘elution plate’.

32. Place ‘elution plate’ into freezer (-20degC) pending processing.

Supplementary Table 10: comparison of PBMC pQTL to whole blood eQTL

‘a32_gtex_v6p_whole_blood_matched_harmonised.tsv’

Comparison of the significant (FDR <0.05) *locally*- acting PBMC pQTL with the *locally*- acting whole blood eQTL from GTEx (v6p) [13].

Column	Description
rsid	rsID of the SNP
chr	Chromosome (GRCh37)
pos	Position (GRCh37)
a1_prot	Effect allele in the protein model
a0_prot	Other allele in the protein model
freq1_prot	Allele frequency of the effect allele
n_missing_geno	Number of individuals with missing data for the SNP in the protein study
n_missing_prot	Number of proteins with missing data for the protein study (of 251)
uniprot_swissprot	UniProtKB / Swiss-Prot ID

ENSG_ID	Ensembl gene ID
gene_name	Ensembl gene name
gene_start	Start position of the gene (Ensembl GRCh37)
gene_end	End position of the gene (Ensembl GRCh37)
tss_distance	Distance to transcription start site (as per GTEx v6p)
beta1_prot	Beta effect estimate (per 'a1_prot') on protein
p_prot	p-value associated with 'beta1_prot'
a1_gtex	Effect allele in GTEx (v6p)
a0_gtex	Other allele in GTEx (v6p)
beta1_gtex	Beta effect estimate (per 'a1_gtex') on expression in GTEx (v6p)
p_nominal_gtex	Nominal p-value of 'beta1_gtex' in GTEx (v6p)
p_nominal_thresh_gtex	Nominal p-value threshold for 'p_nominal_gtex' in GTEx (v6p)
beta1_gtex_harmonised_to_a1_prot	Beta effect estimate (per 'a1_prot') on expression in GTEx (v6p)

Appendix 3: Materials relating to Chapter 4

Supplementary Table 11: Regression results

Significant results from the regression on dominance and absolute additive effect-sizes (Equation 4.4 and Equation 4.5) in the primary set. All significant (Bonferroni correction) results from the primary set presented. The results of these annotations, when analysed in the secondary (replication) set are also included. Bonferroni correction for multiple testing (primary set): CODEX p-value <0.05/96; Encode Segmentation 0.05/14; Cistrome Epigenome 0.05/21.

Bonferroni correction for multiple testing (secondary set), dependent on the number of significant results in the primary set. Dominance: CODEX p-value <0.05/36; Encode Segmentation 0.05/4; Cistrome Epigenome 0.05/9. Additive: CODEX p-value <0.05/1; Encode Segmentation 0.05/1; Cistrome Epigenome 0.05/1.

† indicates successful replication.

All significant loci are with reference to the CpG location. None of the annotations were significant with respect to the SNP location.

Columns headings. ‘Dependent variable’: which model is the result in reference to? ‘Dom’ is Equation 4.4, ‘Add’ is Equation 4.5; ‘Collection’, which collection is the annotation from?;

'Set': the set (primary n = 4,101, secondary n=1,000) the result is from; 'Antibody': the antibody used for the Chip-Seq experiment; 'Beta': the effect-size estimate of the annotation on the dependent variable (as part of the model in Equation 4.4 / Equation 4.5); 'se': the standard error of 'Beta'; 'p-value' the p-value corresponding to 'Beta' and 'se'.

Dependent variable	Collection	Set	Antibody	Beta	se	p-value
Add	CODEX	Primary	KDM5B	0.05	0.01	6.61E-05
Add	CODEX	Secondary	KDM5B	0.02	0.02	1.38E-01
Dom	CODEX	Primary	RUNX3	0.04	0.01	2.50E-11
Dom	CODEX	Secondary	RUNX3	0.05	0.01	4.30E-05†
Dom	CODEX	Primary	RUNX1T1	0.03	0.01	2.76E-04
Dom	CODEX	Secondary	RUNX1T1	0.02	0.02	3.00E-01
Dom	CODEX	Primary	TCF12	0.04	0.01	4.44E-10
Dom	CODEX	Secondary	TCF12	0.05	0.01	2.87E-04†
Dom	CODEX	Primary	TCF3	0.03	0.01	1.79E-04
Dom	CODEX	Secondary	TCF3	0.07	0.02	2.26E-04†
Dom	CODEX	Primary	LMO2	0.04	0.01	5.35E-05
Dom	CODEX	Secondary	LMO2	0.04	0.02	6.57E-02
Dom	CODEX	Primary	BCL6	0.04	0.01	3.66E-08

Dom	CODEX	Secondary	BCL6	0.04	0.02	1.08E-02
Dom	CODEX	Primary	ERG	0.05	0.01	2.46E-12
Dom	CODEX	Secondary	ERG	0.06	0.01	2.63E-05†
Dom	CODEX	Primary	FLI1	0.03	0.01	4.20E-04
Dom	CODEX	Secondary	FLI1	0.04	0.02	1.57E-02
Dom	CODEX	Primary	TAL1	0.03	0.01	1.08E-05
Dom	CODEX	Secondary	TAL1	0.04	0.02	1.21E-02
Dom	CODEX	Primary	GATA2	0.04	0.01	4.47E-06
Dom	CODEX	Secondary	GATA2	0.03	0.02	9.19E-02
Dom	CODEX	Primary	RUNX1	0.04	0.01	1.19E-12
Dom	CODEX	Secondary	RUNX1	0.05	0.01	2.22E-05†
Dom	CODEX	Primary	CBFB	0.03	0.01	6.88E-07
Dom	CODEX	Secondary	CBFB	0.05	0.01	1.07E-04†
Dom	CODEX	Primary	MYH11	0.04	0.01	4.29E-09
Dom	CODEX	Secondary	MYH11	0.05	0.01	2.60E-04†
Dom	CODEX	Primary	MED21	0.04	0.01	5.56E-08
Dom	CODEX	Secondary	MED21	0.06	0.02	9.88E-05†
Dom	CODEX	Primary	TBP	0.03	0.01	1.89E-05
Dom	CODEX	Secondary	TBP	0.04	0.02	1.96E-02
Dom	CODEX	Primary	ELF1	0.03	0.01	3.24E-05
Dom	CODEX	Secondary	ELF1	0.03	0.02	5.35E-02

Dom	CODEX	Primary	SPI1	0.03	0.01	1.06E-04
Dom	CODEX	Secondary	SPI1	0.02	0.01	1.60E-01
Dom	CODEX	Primary	EP300	0.04	0.01	2.47E-07
Dom	CODEX	Secondary	EP300	0.05	0.01	1.07E-03†
Dom	CODEX	Primary	HDAC1	0.03	0.01	4.59E-06
Dom	CODEX	Secondary	HDAC1	0.05	0.01	1.36E-03†
Dom	CODEX	Primary	FOXA2	0.03	0.01	5.79E-07
Dom	CODEX	Secondary	FOXA2	0.03	0.01	1.81E-02
Dom	CODEX	Primary	POLR2A	0.04	0.01	2.77E-13
Dom	CODEX	Secondary	POLR2A	0.06	0.01	3.40E-06†
Dom	CODEX	Primary	ETS1	0.04	0.01	4.94E-10
Dom	CODEX	Secondary	ETS1	0.05	0.01	1.75E-04†
Dom	CODEX	Primary	NOTCH1	0.04	0.01	3.87E-09
Dom	CODEX	Secondary	NOTCH1	0.05	0.01	2.08E-04†
Dom	CODEX	Primary	RBPJ	0.04	0.01	7.33E-06
Dom	CODEX	Secondary	RBPJ	0.05	0.02	2.87E-03
Dom	CODEX	Primary	BRD4	0.03	0.01	1.39E-04
Dom	CODEX	Secondary	BRD4	0.04	0.02	1.84E-02
Dom	CODEX	Primary	MED1	0.03	0.01	1.92E-04
Dom	CODEX	Secondary	MED1	0.03	0.01	1.71E-02
Dom	CODEX	Primary	P300	0.03	0.01	5.02E-05

Dom	CODEX	Secondary	P300	0.04	0.02	1.72E-02
Dom	CODEX	Primary	GATA1	0.03	0.01	4.51E-05
Dom	CODEX	Secondary	GATA1	0.03	0.02	3.93E-02
Dom	CODEX	Primary	CDK7	0.04	0.01	1.54E-06
Dom	CODEX	Secondary	CDK7	0.05	0.02	4.44E-04†
Dom	CODEX	Primary	MYB	0.04	0.01	3.35E-09
Dom	CODEX	Secondary	MYB	0.05	0.01	2.61E-04†
Dom	CODEX	Primary	PHF8	0.04	0.01	1.71E-11
Dom	CODEX	Secondary	PHF8	0.05	0.01	4.50E-04†
Dom	CODEX	Primary	KDM5B	0.03	0.01	8.14E-06
Dom	CODEX	Secondary	KDM5B	0.04	0.01	3.34E-03
Dom	CODEX	Primary	RBBP5	0.04	0.01	1.46E-08
Dom	CODEX	Secondary	RBBP5	0.04	0.01	6.55E-03
Dom	CODEX	Primary	SAP30	0.04	0.01	3.18E-07
Dom	CODEX	Secondary	SAP30	0.03	0.02	3.55E-02
Dom	CODEX	Primary	KDM4A	0.04	0.01	1.34E-08
Dom	CODEX	Secondary	KDM4A	0.04	0.01	4.80E-03
Dom	CODEX	Primary	BCOR	0.04	0.01	5.69E-08
Dom	CODEX	Secondary	BCOR	0.03	0.01	4.29E-02
Add	Encode Segmentation	Primary	Repressed Segments	-0.03	0.01	2.25E-04

Add	Encode Segmentation	Secondary	Repressed Segments	-0.02	0.01	8.62E-02
Dom	Encode Segmentation	Primary	Enhancers Segments	0.02	0.01	3.24E-04
Dom	Encode Segmentation	Secondary	Enhancers Segments	0.02	0.01	8.09E-02
Dom	Encode Segmentation	Primary	Repressed Segments	-0.03	0.00	3.01E-09
Dom	Encode Segmentation	Secondary	Repressed Segments	-0.03	0.01	3.74E-03†
Dom	Encode Segmentation	Primary	Transcribed Segments	-0.04	0.00	6.35E-15
Dom	Encode Segmentation	Secondary	Transcribed Segments	-0.06	0.01	6.20E-08†
Dom	Encode Segmentation	Primary	TSS Segments	0.07	0.01	< 2e-16
Dom	Encode Segmentation	Secondary	TSS Segments	0.11	0.01	< 2e-16†
Add	Cistrome Epigenome	Primary	H3K9K14ac	0.03	0.01	1.48E-04
Add	Cistrome Epigenome	Secondary	H3K9K14ac	0.04	0.01	6.25E-03†
Dom	Cistrome Epigenome	Primary	H3K4me2	0.05	0.00	< 2e-16

Dom	Cistrome Epigenome	Secondary	H3K4me2	0.07	0.01	2.11E-11†
Dom	Cistrome Epigenome	Primary	H3K4me3	0.07	0.01	< 2e-16
Dom	Cistrome Epigenome	Secondary	H3K4me3	0.10	0.01	< 2e-16†
Dom	Cistrome Epigenome	Primary	H3K9ac	0.06	0.01	< 2e-16
Dom	Cistrome Epigenome	Secondary	H3K9ac	0.09	0.01	1.16E-13†
Dom	Cistrome Epigenome	Primary	H3K9K14ac	0.05	0.00	< 2e-16
Dom	Cistrome Epigenome	Secondary	H3K9K14ac	0.06	0.01	2.17E-09†
Dom	Cistrome Epigenome	Primary	H3K27me3	0.02	0.01	2.18E-03
Dom	Cistrome Epigenome	Secondary	H3K27me3	0.02	0.01	1.87E-01
Dom	Cistrome Epigenome	Primary	H3K9me3	0.04	0.00	2.19E-14
Dom	Cistrome Epigenome	Secondary	H3K9me3	0.04	0.01	6.74E-05†
Dom	Cistrome Epigenome	Primary	H3K36me3	-0.03	0.01	1.85E-03

Dom	Cistrome Epigenome	Secondary	H3K36me3	-0.05	0.02	1.22E-02
Dom	Cistrome Epigenome	Primary	Ace_H3	0.05	0.01	< 2e-16
Dom	Cistrome Epigenome	Secondary	Ace_H3	0.07	0.01	3.18E-09†
Dom	Cistrome Epigenome	Primary	H3K14ac	0.04	0.01	1.35E-11
Dom	Cistrome Epigenome	Secondary	H3K14ac	0.05	0.01	1.63E-05†

7. References

1. Chang, E. M., Bretherick, A., Drummond, G. B. & Baillie, J. K. Predictive validity of a novel non-invasive estimation of effective shunt fraction in critically ill patients. *Intensive Care Med. Exp.* **7**, 49 (2019).

2. Zeng, Y. *et al.* Parent of origin genetic effects on methylation in humans are common and influence complex trait variation. *Nat. Commun.* **10**, 1383 (2019).

3. Timmers, P. R. *et al.* Genomics of 1 million parent lifespans implicates novel pathways and common diseases and distinguishes survival chances. *eLife* **8**, (2019).

4. Baillie, J. K. *et al.* Shared activity patterns arising at genetic susceptibility loci reveal underlying genomic and cellular architecture of human disease. *PLoS Comput. Biol.* **14**, e1005934 (2018).

5. Bretherick, A. D. *et al.* Proteome-by-phenome Mendelian Randomisation detects 38 proteins with causal roles in human diseases and traits. *bioRxiv* 631747 (2019) doi:10.1101/631747.

6. Langdon, R. *et al.* Identifying epigenetic biomarkers of established prognostic factors and survival in a clinical cohort of individuals with oropharyngeal cancer. *bioRxiv* 679316 (2019) doi:10.1101/679316.

7. Walker, R. M. *et al.* Assessment of DNA methylation differences between carriers of APOE ε4 and APOE ε2. *bioRxiv* 815035 (2019) doi:10.1101/815035.

8. GoDMC. http://www.godmc.org.uk/.

9. SCALLOP - genetic regulation of the proteome. http://www.scallop-consortium.com/.

10. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *Nat. Genet.* **50**, 1593–1599 (2018).

11. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337**, 1190–1195 (2012).

12. McRae, A. F. *et al.* Identification of 55,000 Replicated DNA Methylation QTL. *Sci. Rep.* **8**, 17605 (2018).

13. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

14. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).

15. Bonder, M. J. *et al.* Disease variants alter transcription factor levels and methylation of their binding sites. *Nat. Genet.* **49**, 131–138 (2017).

16. Ezkurdia, I. *et al.* Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Hum. Mol. Genet.* **23**, 5866–5878 (2014).

17. Clamp, M. *et al.* Distinguishing protein-coding and noncoding genes in the human genome. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 19428–19433 (2007).

18. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **431**, 931–945 (2004).

19. Mirauta, B. A. *et al.* Population-scale proteome variation in human induced pluripotent stem cells. *bioRxiv* (2018) doi:10.1101/439216.

20. Hannon, E. *et al.* Methylation QTLs in the developing brain and their enrichment in schizophrenia risk loci. *Nat. Neurosci.* **19**, 48–54 (2016).

21. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic variation located at thousands of independent loci. *Nat. Commun.* **6**, 6326 (2015).

22. Saxonov, S., Berg, P. & Brutlag, D. L. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl. Acad. Sci. U. S. A.* **103**, 1412–1417 (2006).

23. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25**, 1010–1022 (2011).

24. IL6R Genetics Consortium Emerging Risk Factors Collaboration. Interleukin-6 receptor pathways in coronary heart disease: a collaborative meta-analysis of 82 studies. *Lancet* **379**, 1205–1213 (2012).

25. Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium. The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214–1224 (2012).

26. Burgess, S. *et al.* Using published data in Mendelian randomization: a blueprint for efficient identification of causal risk factors. *Eur. J. Epidemiol.* **30**, 543–552 (2015).

27. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat. Genet.* **48**, 481–487 (2016).

28. Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nat. Genet.* **47**, 856–860 (2015).

29. Canela-Xandri, O., Rawlik, K. & Tenesa, A. An atlas of genetic associations in UK Biobank. *bioRxiv* (2017) doi:10.1101/176834.

30. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

31. Quenneville, S. *et al.* The KRAB-ZFP/KAP1 system contributes to the early embryonic establishment of site-specific DNA methylation patterns maintained during development. *Cell Rep.* **2**, 766–773 (2012).

32. Groner, A. C. *et al.* KRAB–Zinc Finger Proteins and KAP1 Can Mediate Long-Range Transcriptional Repression through Heterochromatin Spreading. *PLoS Genet.* **6**, (2010).

33. Munos, B. Lessons from 60 years of pharmaceutical innovation. *Nat. Rev. Drug Discov.* **8**, 959–968 (2009).

34. Arrowsmith, J. Trial watch: Phase II failures: 2008-2010. *Nat. Rev. Drug Discov.* **10**, 328–329 (2011).

35. Baillie, J. K. Translational genomics. Targeting the host immune response to fight infection. *Science* **344**, 807–808 (2014).

36. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci. Transl. Med.* **9**, eaag1166 (2017).

37. Fang, H. *et al.* A genetics-led approach defines the drug target landscape of 30 immune-related traits. *Nat. Genet.* **51**, 1082–1091 (2019).

38. Behan, F. M. *et al.* Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens. *Nature* **568**, 511–516 (2019).

39. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

40. Smith, G. D. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).

41. Folkersen, L. *et al.* Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease. *PLoS Genet.* **13**, e1006706 (2017).

42. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).

43. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).

44. Zheng, J. *et al.* Phenome-wide Mendelian randomization mapping the influence of the plasma proteome on complex diseases. *bioRxiv* (2019) doi:10.1101/627398.

45. Chong, M. *et al.* Novel Drug Targets for Ischemic Stroke Identified Through Mendelian Randomization Analysis of the Blood Proteome. *Circulation* **140**, 819–830 (2019).

46. Mosley, J. D. *et al.* Probing the Virtual Proteome to Identify Novel Disease Biomarkers. *Circulation* **138**, 2469–2481 (2018).

47. The CARDIoGRAMplusC4D Consortium. A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.* **47**, 1121–1130 (2015).

48. Scott, R. A. *et al.* An expanded genome-wide association study of type 2 diabetes in Europeans. *Diabetes* **66**, 2888–2902 (2017).

49. Nelson, C. P. *et al.* Association analyses based on false discovery rate implicate new loci for coronary artery disease. *Nat. Genet.* **49**, 1385–1391 (2017).

50. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

51. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).

52. Bronson, P. G. *et al.* Common variants at PVT1, ATG13-AMBRA1, AHI1 and CLEC16A are associated with selective IgA deficiency. *Nat. Genet.* **48**, 1425–1429 (2016).

53. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).

54. van Rheenen, W. *et al.* Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. *Nat. Genet.* **48**, 1043–1048 (2016).

55. Hammerschlag, A. R. *et al.* Genome-wide association analysis of insomnia complaints identifies risk genes and genetic overlap with psychiatric and metabolic traits. *Nat. Genet.* **49**, 1584–1592 (2017).

56. Sniekers, S. *et al.* Genome-wide association meta-analysis of 78,308 individuals identifies new loci and genes influencing human intelligence. *Nat. Genet.* **49**, 1107–1112 (2017).

57. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).

58. Hou, L. *et al.* Genome-wide association study of 40,000 individuals identifies two novel loci associated with bipolar disorder. *Hum. Mol. Genet.* **25**, 3383–3394 (2016).

59. Beaumont, R. N. *et al.* Genome-wide association study of offspring birth weight in 86 577 women identifies five novel loci and highlights maternal genetic effects that are independent of fetal genetics. *Hum. Mol. Genet.* **27**, 742–756 (2018).

60. Phelan, C. M. *et al.* Identification of 12 new susceptibility loci for different histotypes of epithelial ovarian cancer. *Nat. Genet.* **49**, 680–691 (2017).

61. van der Harst, P. & Verweij, N. Identification of 64 Novel Genetic Loci Provides an Expanded View on the Genetic Architecture of Coronary Artery Disease. *Circ. Res.* **122**, 433–443 (2018).

62. van den Berg, S. M. *et al.* Meta-analysis of genome-wide association studies for extraversion: findings from the Genetics of Personality Consortium. *Behav. Genet.* **46**, 170–182 (2016).

63. Genetics of Personality Consortium. Meta-analysis of genome-wide association studies for neuroticism, and the polygenic association with major depressive disorder. *JAMA Psychiatry* **72**, 642–650 (2015).

64. The EArly Genetics and Lifecourse Epidemiology (EAGLE) Eczema Consortium. Multi-ancestry genome-wide association study of 21,000 cases and 95,000 controls identifies new risk loci for atopic dermatitis. *Nat. Genet.* **47**, 1449–1456 (2015).

65. Ferreira, M. A. *et al.* Shared genetic origin of asthma, hay fever and eczema elucidates allergic disease biology. *Nat. Genet.* **49**, 1752–1757 (2017).

66. Astle, W. J. *et al.* The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415-1429.e19 (2016).

67. Staley, J. R. *et al.* PhenoScanner: a database of human genotype–phenotype associations. *Bioinformatics* **32**, 3207–3209 (2016).

68. Kamat, M. A. *et al.* PhenoScanner V2: an expanded tool for searching human genotype-phenotype associations. *Bioinformatics* **35**, 4851–4853 (2019).

69. Gordon, E. D. *et al.* IL1RL1 asthma risk variants regulate airway type 2 inflammation. *JCI Insight* **1**, e87871 (2016).

70. Gudbjartsson, D. F. *et al.* Sequence variants affecting eosinophil numbers associate with asthma and myocardial infarction. *Nat. Genet.* **41**, 342–347 (2009).

71. Busse, W. W. *et al.* Daclizumab improves asthma control in patients with moderate to severe persistent asthma: a randomized, controlled trial. *Am. J. Respir. Crit. Care Med.* **178**, 1002–1008 (2008).

72. Massoud, A. H. *et al.* An asthma-associated IL4R variant exacerbates airway inflammation by promoting conversion of regulatory T cells to TH17-like cells. *Nat. Med.* **22**, 1013–1022 (2016).

73. Navarini, A. A., French, L. E. & Hofbauer, G. F. L. Interrupting IL-6–receptor signaling improves atopic dermatitis but associates with bacterial superinfection. *J. Allergy Clin. Immunol.* **128**, 1128–1130 (2011).

74. Ullah, M. A., Sukkar, M., Ferreira, M. & Phipps, S. 53: IL-6R blockade: A new personalised treatment for asthma? *Cytokine* **70**, 40 (2014).

75. Esparza-Gordillo, J. *et al.* A functional IL-6 receptor (IL6R) variant is a risk factor for persistent atopic dermatitis. *J. Allergy Clin. Immunol.* **132**, 371–377 (2013).

76. Ferreira, M. A. R. *et al.* Identification of IL6R and chromosome 11q13.5 as risk loci for asthma. *Lancet* **378**, 1006–1014 (2011).

77. Scott, L. J. Tocilizumab: a review in rheumatoid arthritis. *Drugs* **77**, 1865–1879 (2017).

78. Thul, P. J. *et al.* A subcellular map of the human proteome. *Science* **356**, eaal3321 (2017).

79. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).

80. Uhlen, M. *et al.* A pathology atlas of the human cancer transcriptome. *Science* **357**, eaan2507 (2017).

81. Ohnishi, H. *et al.* Differential localization of Src homology 2 domain-dontaining protein tyrosine phosphatase substrate-1 and CD47 and Its molecular mechanisms in cultured hippocampal neurons. *J. Neurosci.* **25**, 2702–2711 (2005).

82. Toth, A. B. *et al.* Synapse maturation by activity-dependent ectodomain shedding of SIRPα. *Nat. Neurosci.* **16**, 1417–1425 (2013).

83. Ma, L., Kulesskaya, N., Võikar, V. & Tian, L. Differential expression of brain immune genes and schizophrenia-related behavior in C57BL/6N and DBA/2J female mice. *Psychiatry Res.* **226**, 211–216 (2015).

84. Koshimizu, H., Takao, K., Matozaki, T., Ohnishi, H. & Miyakawa, T. Comprehensive behavioral analysis of cluster of differentiation 47 knockout mice. *PLoS ONE* **9**, e89584 (2014).

85. Ohnishi, H. *et al.* Stress-evoked tyrosine phosphorylation of signal regulatory protein α regulates behavioral immobility in the forced swim test. *J. Neurosci.* **30**, 10472–10483 (2010).

86. Chang, H. P., Lindberg, F. P., Wang, H. L., Huang, A. M. & Lee, E. H. Y. Impaired memory retention and decreased long-term potentiation in integrin-associated protein-deficient mice. *Learn. Mem.* **6**, 448–457 (1999).

87. Huang, A. M., Wang, H. L., Tang, Y. P. & Lee, E. H. Y. Expression of integrin-associated protein gene associated with memory formation in rats. *J. Neurosci.* **18**, 4305–4313 (1998).

88. Brown, G. C. & Neher, J. J. Microglial phagocytosis of live neurons. *Nat. Rev. Neurosci.* **15**, 209–216 (2014).

89. Martins-de-Souza, D. *et al.* Prefrontal cortex shotgun proteome analysis reveals altered calcium homeostasis and immune system imbalance in schizophrenia. *Eur. Arch. Psychiatry Clin. Neurosci.* **259**, 151–163 (2009).

90. Klarin, D. *et al.* Genetic analysis in UK Biobank links insulin resistance and transendothelial migration pathways to coronary artery disease. *Nat. Genet.* **49**, 1392–1397 (2017).

91. Ozaki, K. *et al.* Functional SNPs in the lymphotoxin-alpha gene that are associated with susceptibility to myocardial infarction. *Nat. Genet.* **32**, 650–654 (2002).

92. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).

93. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

94. Ferreira, R. C. *et al.* Functional IL6R 358Ala allele impairs classical IL-6 receptor signaling and influences risk of diverse inflammatory diseases. *PLoS Genet.* **9**, e1003444 (2013).

95. Wenzel, S. *et al.* Dupilumab efficacy and safety in adults with uncontrolled persistent asthma despite use of medium-to-high-dose inhaled corticosteroids plus a long-acting β2 agonist: a randomised double-blind placebo-controlled pivotal phase 2b dose-ranging trial. *Lancet* **388**, 31–44 (2016).

96. Wenzel, S. *et al.* Dupilumab in persistent asthma with elevated eosinophil levels. *N. Engl. J. Med.* **368**, 2455–2466 (2013).

97. McQuillan, R. *et al.* Runs of homozygosity in European populations. *Am. J. Hum. Genet.* **83**, 359–372 (2008).

98. Campbell, H. *et al.* Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum. Mol. Genet.* **16**, 233–241 (2007).

99. Rudan, I. *et al.* '10001 Dalmatians:' Croatia launches its national biobank. *Croat. Med. J.* **50**, 4–6 (2009).

100. Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).

101. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, s13742-015-0047–8 (2015).

102. Purcell, S. *PLINK: v1.90*. (2017).

103. O'Connell, J. *et al.* A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet.* **10**, e1004234 (2014).

104. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).

105. Assarsson, E. *et al.* Homogenous 96-plex PEA immunoassay exhibiting high sensitivity, specificity, and excellent scalability. *PloS ONE* **9**, e95192 (2014).

106. Haller, T., Kals, M., Esko, T., Mägi, R. & Fischer, K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief. Bioinform.* **16**, 39–44 (2015).

107. Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res.* **46**, D754–D761 (2018).

108. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits*. (Sinauer, 1998).

109. Davies, M. *et al.* ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.* **43**, W612–W620 (2015).

110. Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337 (2010).

111. Končarević, S. *et al.* In-Depth Profiling of the Peripheral Blood Mononuclear Cells Proteome for Clinical Blood Proteomics. *International Journal of Proteomics* (2014) doi:10.1155/2014/129259.

112. Rohloff, J. C. *et al.* Nucleic Acid Ligands With Protein-like Side Chains: Modified Aptamers and Their Use as Diagnostic and Therapeutic Agents. *Mol. Ther. - Nucleic Acids* **3**, e201 (2014).

113. SomaLogic. *Short-Technical-Note-SOMAmer-specificity.pdf*. (2019).

114. Smith, B. H. *et al.* Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int. J. Epidemiol.* **42**, 689–700 (2013).

115. Smith, B. H. *et al.* Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med. Genet.* **7**, 74 (2006).

116. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).

117. Yang, J., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: a tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).

118. Ishihama, Y., Rappsilber, J. & Mann, M. Modular stop and go extraction tips with stacked disks for parallel and multidimensional Peptide fractionation in proteomics. *J. Proteome Res.* **5**, 988–994 (2006).

119. Rappsilber, J., Mann, M. & Ishihama, Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906 (2007).

120. UniProt Knowledgebase (UniProtKB) Release 2019_01. https://www.uniprot.org/ (2019).

121. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics MCP* **13**, 2513–2526 (2014).

122. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).

123. Utz, P. J., Genovese, M. C. & Robinson, W. H. Unlocking the "PAD" lock on rheumatoid arthritis. *Ann. Rheum. Dis.* **63**, 330–332 (2004).

124. Spengler, J. *et al.* Release of Active Peptidyl Arginine Deiminases by Neutrophils Can Explain Production of Extracellular Citrullinated Autoantigens in Rheumatoid Arthritis Synovial Fluid. *Arthritis Rheumatol. Hoboken Nj* **67**, 3135–3145 (2015).

125. Kinloch, A. *et al.* Synovial fluid is a site of citrullination of autoantigens in inflammatory arthritis. *Arthritis Rheum.* **58**, 2287–2295 (2008).

126. Willis, V. C. *et al.* N-α-Benzoyl-N5-(2-Chloro-1-Iminoethyl)-L-Ornithine Amide, a Protein Arginine Deiminase Inhibitor, Reduces the Severity of Murine Collagen-Induced Arthritis. *J. Immunol. Baltim. Md 1950* **186**, 4396–4404 (2011).

127. Kawalkowska, J. *et al.* Abrogation of collagen-induced arthritis by a peptidyl arginine deiminase inhibitor is associated with modulation of T cell-mediated immune responses. *Sci. Rep.* **6**, (2016).

128. Willis, V. C. *et al.* Protein arginine deiminase 4 inhibition is sufficient for the amelioration of collagen-induced arthritis. *Clin. Exp. Immunol.* **188**, 263–274 (2017).

129. Fuhrmann, J. & Thompson, P. R. Protein Arginine Methylation and Citrullination in Epigenetic Regulation. *ACS Chem. Biol.* **11**, 654–668 (2016).

130. Zhang, X. *et al.* Peptidylarginine deiminase 2-catalyzed histone H3 arginine 26 citrullination facilitates estrogen receptor α target gene activation. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 13331–13336 (2012).

131. Hao, G.-F., Li, Y.-S., Liu, J.-L. & Wo, M.-Y. Association of HLA-DQA1 (rs9272219) with susceptibility to rheumatoid arthritis in a Han Chinese population. *Int. J. Clin. Exp. Pathol.* **7**, 8155–8158 (2014).

132. Eleftherohorinou, H., Hoggart, C. J., Wright, V. J., Levin, M. & Coin, L. J. M. Pathway-driven gene stability selection of two rheumatoid arthritis GWAS identifies and validates new susceptibility genes in receptor mediated signalling pathways. *Hum. Mol. Genet.* **20**, 3494–3506 (2011).

133. Guo, J. *et al.* Sequencing of the MHC region defines HLA-DQA1 as the major genetic risk for seropositive rheumatoid arthritis in Han Chinese population. *Ann. Rheum. Dis.* **78**, 773–780 (2019).

134. Han, B. *et al.* Fine Mapping Seronegative and Seropositive Rheumatoid Arthritis to Shared and Distinct HLA Alleles by Adjusting for the Effects of Heterogeneity. *Am. J. Hum. Genet.* **94**, 522–532 (2014).

135. Raychaudhuri, S. *et al.* Five amino acids in three HLA proteins explain most of the association between MHC and seropositive rheumatoid arthritis. *Nat. Genet.* **44**, 291–296 (2012).

136. Afroz, S. *et al.* A Comprehensive Gene Expression Meta-analysis Identifies Novel Immune Signatures in Rheumatoid Arthritis Patients. *Front. Immunol.* **8**, 74 (2017).

137. Orozco, G. *et al.* HLA-DPB1-COL11A2 and three additional xMHC loci are independently associated with RA in a UK cohort. *Genes Immun.* **12**, 169–175 (2011).

138. Zhu, H. *et al.* Gene-Based Genome-Wide Association Analysis in European and Asian Populations Identified Novel Genes for Rheumatoid Arthritis. *PLOS ONE* **11**, e0167212 (2016).

139. Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nat. Genet.* **51**, 768 (2019).

140. Lea, A. J. *et al.* Genome-wide quantification of the effects of DNA methylation on human gene regulation. *eLife* **7**, e37513 (2018).

141. Yin, Y. *et al.* Impact of cytosine methylation on DNA binding specificities of human transcription factors. *Science* **356**, (2017).

142. Wang, H., Lou, D. & Wang, Z. Crosstalk of Genetic Variants, Allele-Specific DNA Methylation, and Environmental Factors for Complex Disease Risk. *Front. Genet.* **9**, (2019).

143. Wong, E. S. *et al.* Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. *Nat. Commun.* **8**, 1092 (2017).

144. Zhu, H., Wang, G. & Qian, J. Transcription factors as readers and effectors of DNA methylation. *Nat. Rev. Genet.* **17**, 551–565 (2016).

145. *Davidson's Principles and Practice of Medicine*. (Elsevier, 2006).

146. Falconer, D. S. *Introduction to Quantitative Genetics*. (Oliver and Boyd, 1960).

147. Nagy, R. *et al.* Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Med.* **9**, 23 (2017).

148. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

149. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler transform (PBWT). *Bioinforma. Oxf. Engl.* **30**, 1266–1272 (2014).

150. Fortin, J.-P., Fertig, E. & Hansen, K. shinyMethyl: interactive quality control of Illumina 450k DNA methylation arrays in R. *F1000Research* **3**, 175 (2014).

151. Patterson, N., Price, A. L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).

152. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

153. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinforma. Oxf. Engl.* **30**, 1363–1369 (2014).

154. Szklarczyk, D. *et al.* STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.* **47**, D607–D613 (2019).

155. Ensembl Biomart GRCh37. http://grch37.ensembl.org/.

156. Gene Ontology Resource. http://geneontology.org/.

157. Reactome Pathway Database. https://reactome.org/.

158. UniProt. https://www.uniprot.org/.

159. Pfam. https://pfam.xfam.org.

160. InterPro. https://www.ebi.ac.uk/interpro/.

161. SMART. http://smart.embl.de/.

162. Ensembl genome browser 98. https://www.ensembl.org.

163. Kent, W. J. *et al.* The Human Genome Browser at UCSC. *Genome Res.* **12**, 996–1006 (2002).

164. UCSC Genome Browser. http://genome.ucsc.edu/.

165. Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinforma. Oxf. Engl.* **32**, 587–589 (2016).

166. Liu, T. *et al.* Cistrome: an integrative platform for transcriptional regulation studies. *Genome Biol.* **12**, R83 (2011).

167. Sánchez-Castillo, M. *et al.* CODEX: a next-generation sequencing experiment database for the haematopoietic and embryonic stem cell communities. *Nucleic Acids Res.* **43**, D1117–D1123 (2015).

168. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

169. Albert, F. W. & Kruglyak, L. The role of regulatory variation in complex traits and disease. *Nat. Rev. Genet.* **16**, 197–212 (2015).

170. Veitia, R. A., Caburet, S. & Birchler, J. A. Mechanisms of Mendelian dominance. *Clin. Genet.* **93**, 419–428 (2018).

171. Omholt, S. W., Plahte, E., Oyehaug, L. & Xiang, K. Gene regulatory networks generating the phenomena of additivity, dominance and epistasis. *Genetics* **155**, 969–980 (2000).

172. Adya, N., Castilla, L. H. & Liu, P. P. Function of CBFβ/Bro proteins. *Semin. Cell Dev. Biol.* **11**, 361–368 (2000).

173. Wilson-Rawls, J., Molkentin, J. D., Black, B. L. & Olson, E. N. Activated notch inhibits myogenic activity of the MADS-Box transcription factor myocyte enhancer factor 2C. *Mol. Cell. Biol.* **19**, 2853–2862 (1999).

174. Imbeault, M., Helleboid, P.-Y. & Trono, D. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* **543**, 550–554 (2017).

175. Ecco, G., Imbeault, M. & Trono, D. KRAB zinc finger proteins. *Dev. Camb. Engl.* **144**, 2719–2729 (2017).

176. Urrutia, R. KRAB-containing zinc-finger repressor proteins. *Genome Biol.* **4**, 231 (2003).

177. Schultz, D. C., Ayyanathan, K., Negorev, D., Maul, G. G. & Rauscher, F. J. SETDB1: a novel KAP-1-associated histone H3, lysine 9-specific methyltransferase that contributes to HP1-mediated silencing of euchromatic genes by KRAB zinc-finger proteins. *Genes Dev.* **16**, 919–932 (2002).

178. Schultz, D. C., Friedman, J. R. & Rauscher, F. J. Targeting histone deacetylase complexes via KRAB-zinc finger proteins: the PHD and bromodomains of KAP-1 form a cooperative unit that recruits a novel isoform of the Mi-2alpha subunit of NuRD. *Genes Dev.* **15**, 428–443 (2001).

179. Nielsen, A. L. *et al.* Interaction with members of the heterochromatin protein 1 (HP1) family and histone deacetylation are differentially involved in transcriptional silencing by members of the TIF1 family. *EMBO J.* **18**, 6385–6395 (1999).

180. Sripathy, S. P., Stevens, J. & Schultz, D. C. The KAP1 corepressor functions to coordinate the assembly of de novo HP1-demarcated microenvironments of heterochromatin required for KRAB zinc finger protein-mediated transcriptional repression. *Mol. Cell. Biol.* **26**, 8623–8638 (2006).

181. Ryan, R. F. *et al.* KAP-1 Corepressor Protein Interacts and Colocalizes with Heterochromatic and Euchromatic HP1 Proteins: a Potential Role for Krüppel-Associated Box–Zinc Finger Proteins in Heterochromatin-Mediated Gene Silencing. *Mol. Cell. Biol.* **19**, 4366–4378 (1999).

182. Lechner, M. S., Begg, G. E., Speicher, D. W. & Rauscher, F. J. Molecular Determinants for Targeting Heterochromatin Protein 1-Mediated Gene Silencing: Direct Chromoshadow Domain–KAP-1 Corepressor Interaction Is Essential. *Mol. Cell. Biol.* **20**, 6449–6465 (2000).

183. Wiznerowicz, M. *et al.* The Kruppel-associated box repressor domain can trigger de novo promoter methylation during mouse early embryogenesis. *J. Biol. Chem.* **282**, 34535–34541 (2007).

184. Barde, I. *et al.* A KRAB/KAP1-miRNA Cascade Regulates Erythropoiesis Through Stage-Specific Control of Mitophagy. *Science* **340**, 350–353 (2013).

185. Gilbert, L. A. *et al.* CRISPR-Mediated Modular RNA-Guided Regulation of Transcription in Eukaryotes. *Cell* **154**, 442–451 (2013).

186. Bartels, S. J. J. *et al.* A SILAC-based screen for Methyl-CpG binding proteins identifies RBP-J as a DNA methylation and sequence-specific binding protein. *PloS One* **6**, e25884 (2011).

187. Kitsera, N. *et al.* Functional impacts of 5-hydroxymethylcytosine, 5-formylcytosine, and 5-carboxycytosine at a single hemi-modified CpG dinucleotide in a gene promoter. *Nucleic Acids Res.* **45**, 11033–11042 (2017).

188. Timpson, N. J. *et al.* C-reactive protein levels and body mass index: elucidating direction of causation through reciprocal Mendelian randomization. *Int. J. Obes. 2005* **35**, 300–308 (2011).

189. Hemani, G., Tilling, K. & Smith, G. D. Orienting the causal relationship between imprecisely measured traits using GWAS summary data. *PLOS Genet.* **13**, e1007081 (2017).